Informedia News-On Demand: Using Speech Recognition to Create a Digital Video Library

Howard D. Wactlar[1], Alexander G. Hauptmann[1]
and Michael J. Witbrock[2,3]
March 19th, 1998

CMU-CS-98-109

[1]School of Computer Science,
Carnegie Mellon University,
Pittsburgh, PA 15213-3890 USA

[2]Justresearch (Justsystem Pittsburgh Research Center),
4616 Henry St,
Pittsburgh PA 15213 USA

This work was first presented at the 1996 DARPA Spoken Language Technology Workshop, Arden House, Harriman, New York, February 1996.

[3]The work described in this paper was done while the third author was an employee of Carnegie Mellon University

Abstract

In theory, speech recognition technology can make any spoken words in video or audio media usable for text indexing, search and retrieval. This article describes the News-on-Demand application created within the Informedia™ Digital Video Library project and discusses how speech recognition is used in transcript creation from video, alignment with closed-captioned transcripts, audio paragraph segmentation and a spoken query interface. Speech recognition accuracy varies dramatically depending on the quality and type of data used. Informal information retrieval tests show that reasonable recall and precision can be obtained with only moderate speech recognition accuracy.

This paper is based on work supported by the National Science Foundation, DARPA and NASA under NSF Cooperative agreement No. IRI-9411299. We thank Justsystem Corporation for supporting the preparation of the paper. The views and conclusions contained in this document are those of the authors and do not necessarily reflect the views of any of the sponsors.

Keywords: Digital Libraries, Digital Video, Speech Recognition, Image Analysis, Video Segmentation, Information Retrieval, Spoken Document Retrieval, Speech Interfaces, Informedia.

Informedia: News-on-Demand

The Informedia™ digital video library project (Informedia 1997, Christel, Stevens & Wactlar 1994, Stevens, Christel & Wactlar 1994, Christel et al. 1994) at Carnegie Mellon University is creating a digital library of text, images, videos and audio data available for full content search and retrieval. News-on-Demand is an application within Informedia that monitors news from TV, radio and text sources and allows the user to retrieve news stories of interest.

A compelling application of the Informedia project is the indexing and retrieval of television, radio and text news. The Informedia: News-on-Demand application (Hauptmann, Witbrock, Rudnicky and Reed 1995, Hauptmann, Witbrock and Christel 1995) is an innovative example of indexing and searching broadcast news video and news radio material by text content. News-on-Demand is a fully-automatic system that monitors TV, radio and text news and allows selective retrieval of news stories based on spoken queries. The user may choose among the retrieved stories and play back news stories of interest. The system runs on a Pentium PC using MPEG-I video compression. Speech recognition is done on a separate platform using the Sphinx-II continuous speech recognition system (CMU Speech 1997).

The News-on-Demand application forces us to consider the limits of what can be done automatically and in limited time. News events happen daily and it is not feasible to process, segment and label news through manual or "human-assisted" methods. Timeliness of the library information is important, as is the ability to continuously update the contents. Thus we are forced to fully exploit the potential of computer speech recognition without the benefit of human corrections and editing.

Even though our work is centered around processing news stories from TV broadcasts, the system exemplifies an approach that can make any video, audio or text data accessible. Similar methods can help to index and search other streamed multi-media data by content in other Informedia applications.

Other attempts at solutions have been obliged to restrict the data to only text material, as found in most news databases. Video-on-demand allows a user to select (and pay for) a complete program, but does not allow selective retrieval. The closest approximation to News-on-Demand can be found in the "CNN-AT-WORK" system offered to businesses by a CNN/Intel venture. At the heart of the CNN-AT-WORK solution is a digitizer that encodes the video into the INDEO compression format and transmits it to workstations over a local area network. Users can store headlines together with video clips and retrieve them at a later date. However, this service depends entirely on the separately transmitted "headlines" and does not include other news sources. In addition, CNN-AT-WORK does not feature an integrated multimodal query interface (CNN 1996).

Preliminary investigations on the use of speech recognition to analyze a news story were made by (Schauble & Wechsler 1995). Without a powerful speech recognizer, their approach used a phonetic engine that transformed the spoken text into an (errorful) phoneme string. The query was also transformed into a phoneme string and the database searched for the best approximate match. Errors in recognition, as well as word prefix and suffix differences did not severely affect the system since they scattered equally over all documents and well-matching search scores dominate the retrieval.

Another news processing systems that includes video materials is the MEDUSA system (Brown et al. 1995). The MEDUSA news broadcast application can digitize and record news video and teletext, which is equivalent to closed-captions. Instead of segmenting the news into stories, the system uses overlapping windows of adjacent text lines for indexing and retrieval. During retrieval the system responds to typed requests returning an ordered list of the most relevant news broadcasts. Query words are stripped of suffixes before search and the relevance ranking takes word frequency in the segment and over all the corpus into account, as well as the ability of words to discriminate between stories. Within a news broadcast, it is up to the user to select and play a region using information given by the system about the location of the matched keywords. The focus of MEDUSA is in the system architecture and the information retrieval component. No image processing and no speech recognition is performed.

Component Technologies

There are three broad categories of technologies we can bring to bear to create and search a digital video library from broadcast video and audio materials (Hauptmann & Smith 1995)

Text processing looks at the textual (ASCII) representation of the words that were spoken, as well as other text annotations. These may be derived from the transcript, from the production notes or from the closed-captioning that might be available. Text analysis can work on an existing transcript to help segment the text into paragraphs (Mauldin 1991). An analysis of keyword prominence allows us to identify important sections in the transcript (Salton & McGill 1983). Other more sophisticated language based criteria are under investigation. We currently use two main techniques for text analysis:

1. If we have a complete time aligned transcript available from the closed-captioning or through a human-generated transcription, we can exploit natural "structural" text markers such as punctuation to identify news story boundaries

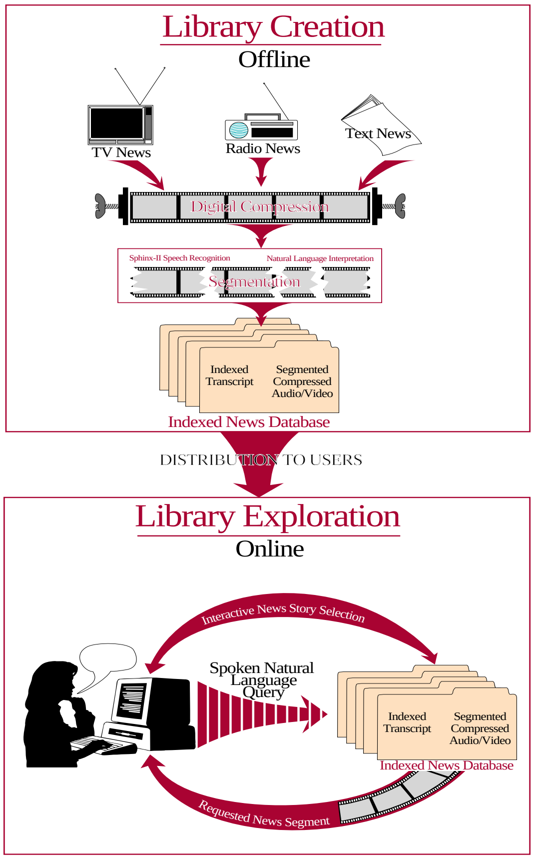

Figure 1. Overview of the News-on-Demand System

2. To identify and rank the contents of one news segment, we use the well-known technique of TF/IDF (term frequency/inverse document frequency) to identify critical keywords and their relative importance for the video document (Salton & McGill 1983).:

Image analysis looks at the images in the video portion of the MPEG stream. This analysis is primarily used for the identification of scene breaks and to select static frame icons that are representative of a scene. Image statistics are computed for primitive image features such as color histograms, and these are used for indexing, matching and segmenting images (Zhang, Low & Smoliar 1995).

Using *color histogram analysis*, video is segmented into scenes through comparative difference measures. Images with little histogram disparity are considered to be relatively equivalent. By detecting significant changes in the weighted color histogram of successive frames, image sequences can be separated into individual scenes. A comparison between the cumulative distributions is used as a difference measure. This result is passed through a high pass filter to further isolate peaks and an empirically derived threshold is used to select only regions where scene breaks occur.

Optical flow analysis is an important method of visual segmentation and description based on interpreting camera motion. We can interpret camera motion as a pan or zoom by examining the geometric properties of the optical flow vectors. Using the Lucas-Kanade gradient descent method for optical flow, we can track individual regions from one frame to the next. By measuring the velocity that individual regions show over time, a motion representation of the scene is created. Drastic changes in this flow indicate random motion, and therefore, new scenes. These changes will also occur during gradual transitions between images such as fades or special effects.

Speech analysis provides the basis for analyzing the audio component of the news broadcast. To transcribe the content of the video material, we use the Sphinx-II system, a large-vocabulary, speaker-independent, continuous speech recognizer created at Carnegie Mellon (CMU Speech 1997, Hwang et al. 1994). Sphinx-II uses senonic semi-continuous hidden Markov models (HMMs) to model between-word context-dependent phones. The system uses four types of codebooks: mel-frequency cepstral coefficients, 1st cepstral differences, 2nd cepstral differences, as well as power and its first and second differences. Twenty-seven phone classes are identified, and a set of four VQ codebooks is trained for each phone class. Cepstral vectors are normalized with an utterance-based cepstral mean value. The semi-continuous observation probability is computed using a variable-sized mixture of the top Gaussian distributions from each phone-dependent codebook. The recognizer processes an utterance in four steps:

1. A forward time-synchronous pass using between-word senonic semi-continuous acoustic models with phone-dependent codebooks and a bigram language model is performed. This produces a set of possible word occurrences, with each word occurrence having one start time and multiple possible end times.

2. A backward pass using the same system configuration is then performed, resulting in multiple possible begin times for each end time predicted in the first pass.

3. An A* algorithm generates the set of N-best hypotheses for the utterance from the results of the forward and backward passes. Any language model can be applied in this pass -- the default is a trigram language model. This approximate A* algorithm is not guaranteed to produce the best-scoring hypothesis first.

4. The best-scoring hypothesis is selected from the N-best list produced. This hypothesis is the recognizer's result.

The language model consists of words with probabilities, bigrams/trigrams which are word pairs/triplets with conditional probabilities for the last word given the previous word(s). Normally, a word trigram is used to predict the next words for the recognizer. However, a backoff procedure allows the next word to be predicted from only the current word (bigram) at a penalty. A word may also occur independently of context based on it's individual probability with another larger backoff penalty. Our current largest and most accurate language model was constructed from a corpus of news stories from the Wall Street Journal from 1989 to 1994 and the Associated Press news service stories from 1988 to 1990. Only trigrams that were encountered more than once were included in the model, but all bigrams and unigrams of the most frequent 58800 words in the corpus (Rudnicky 1995) are used.

Library Creation

Library creation deals with the accumulation of information, transcription, segmentation and indexing. Unlike other Informedia prototypes (Christel, Stevens & Wactlar 1994, Christel et al. 1994) which are designed for educational uses, News-on-Demand focuses on finding rapid and fully automatic methods for the creation of an indexed digital video library. While the long-lived educational content of the other Informedia prototypes allows time for careful computer-assisted human editing, the frequently short-lived value of news requires library creation and update for News-on-Demand to be completely automatic. In early work on the Informedia digital video library, all segmentation was done by hand. Due to the time constraints imposed by continuous daily news coverage and due to the volume of data, we are now using fully automatic news library creation methods. The following steps are performed during library creation by a coordinated suite of programs:

1. **Digitize the video and audio data into MPEG-I compression format.** Our current library creation process starts with a raw digitized video tape or live broadcast. Generally, the videos in the Informedia: News-on-Demand Library are half-hour evening news broadcasts. Radio news shows such as "All Things Considered" or the news broadcasts from the Voice of America are also compressed and analyzed into the library. Using relatively inexpensive off-the-shelf PC-based hardware we can compress video and audio to about 520 MB/hour of video in MPEG-I format. The audio component of the MPEG stream is compressed to about 80 MB/hour.

2. **Create a time-aligned transcript (from closed-captioning and/or speech recognition).** The audio portion is fed through the speech analysis routines, which produces a transcript of the spoken text. A number of the news stories also have closed-captioning text available. However, the closed-captioned data, if available, may lag up to 25 seconds behind the actual words spoken. These timing problems, and inaccuracies in transcription are especially glaring when the broadcast is ``live". In News-on-Demand we use speech recognition in conjunction with closed-captioning, when available, to improve the time-alignment of the transcripts. To create a time-aligned transcript from closed-captioned text, the speech recognizer output is aligned against the closed-caption transcript words using a standard dynamic programming algorithm based on orthographic match. Through this alignment each word in the closed-captioned transcript is assigned a time marker derived from the corresponding word in the speech recognition output. For the news broadcasts that are not closed-captioned, we use a transcript generated exclusively by the speech recognition system. The vocabulary and language model used here approximate a "general American news" language model. These were based on a large corpus of North American business news from 1987 to 1994 (Rudnicky 1995). In addition, transcripts or closed-captioning text maybe obtainable for the video data. If there is a text transcript available during library creation, speech recognition helps create a time-aligned transcript of the spoken words and aids precise segmentation of the broadcast into paragraphs. The library index needs very specific information about the start and end of each spoken word, in order to select the relevant video "paragraph" to retrieve and to find individual words within the story.

3. **Segment story boundaries.** To allow efficient access to the relevant content of the news data, we need to break up the broadcasts into small pieces or into news stories. To answer a user query by showing a half-hour long news show is rarely a reasonable response.The speech signal is also analyzed for low energy sections that indicate acoustic "paragraph" breaks through silence.This is the first pass at segmentation. If a closed-captioned text transcript is available, we use structural markers such as punctuation and paragraph boundaries to identify news stories. If only a speech recognition generated transcript is available, we use acoustically determined paragraph boundaries of silence near extrema of 30 second intervals.

4. **Segment images by scene breaks and select key frames.** Image analysis is primarily used for the identification of breaks between scenes and the identification of a single static frame icon that is representative of a scene. Image statistics methods are used to describe primitive image features such as

color histograms, and their time functions. These are used for indexing, matching and segmenting images.

5. **Index all stories.** The keywords and their corresponding paragraph locations in the video are indexed in the informedia library catalogue. An inverted index is created using a modified version of the Pursuit search engine (Mauldin 1991). To obtain video clips suitable for viewers, we first search for keywords from the user query in the recognition transcript. When we find a match, the surrounding video paragraph is returned using the timing information derived from speech recognition.

Library Exploration

Library exploration is the interaction between the system and the user trying to retrieve selections in the database. Users are able to explore the Informedia library through an interface that allows them to search using typed or spoken natural language queries, review and select relevant documents retrieved from the library and play/display the material on their PC workstations. During library exploration the user generally goes through the following procedure, where steps may be repeated or skipped:

1. **Speaking a query.** During library exploration, the Sphinx-II (Hwang et al. 1994) speech recognition system allows a user to query Informedia by voice, simplifying the interface by making the interaction more direct. A query text window shows the result of the speech recognition on the spoken query and can be edited by selecting portions of the text and typing. The language model used during exploration is similar to that in (Rudnicky 1995) but emphasizes key phrases frequently used in queries such as "How about", "Tell me about", etc. In the future we plan to limit the language model to reflect those words found in the actual library data, making the language model more efficient and smaller.

2. **Submitting a query to search the database.** Stopwords are eliminated from the query. The stopwords are function words from among the most frequent query words. The query also does term expansion looking for words with the same stems derived from a Houghton-Mifflin electronic dictionary of word roots. Retrieval is done through an inverted index, and each indexed story is ranked by relevance based on the frequency of the keywords occurring within the news story relative to the frequency of the keyword in all news stories.

3. **Selection among the returned query results.** The results of a query are displayed through keyframe icons. The most relevant words from each story are displayed at the top of each icon when the mouse passes over it. To find out more about the video story the user can click on the filmstrip button. In the filmstrip view, one keyframe thumbnail icon is displayed for each scene in the news story. Moving the mouse over the scene displays the keywords that were matched within the scene. The location of these words is also precisely marked. These features help the user choose effectively among the returned results.

4. **Playing or displaying a story**. Once the user has decided and clicked on a story icon to play, the matching news story is played in the video window Transcript text can be displayed below the video window and scrolls together with the video.

News-on-Demand Preliminary Speech Recognition Findings

Table 1 shows the results from informal recognition experiments with different video data. The results on our

Type of Speech Data	**Word Error Rate =** Insertion+Deletion+Substitution
1) Speech benchmarks evaluation	~ 8% - 12%
2) News text spoken in lab	~ 10%- 17%
3) Narrator recorded in TV studio	~ 20%
4) C-Span	~ 40%
5) Dialog in documentary video	~ 50% - 65%
6) Evening News (30 min)	~ 55%
7) Complete documentary	~ 75%
8) Commercials	~ 85%

Table 1: Preliminary Speech Recognition Results

data confirm that the type of data and the environment in which it was created dramatically alters the speech recognition accuracy.

1. The basic reference point is the standard speech evaluation data which is used to benchmark speech recognition systems with large vocabularies between five thousand and sixty thousand words. The recog-

nition systems are carefully tuned to this evaluation and the results can be considered close to optimal for the current state of speech recognition research. In these evaluations, we typically see word error rates ranging from 8 percent to 12 percent depending on the test set. Note that word error rate is defined as the sum of insertions, substitutions and deletions. This value can be larger than 100 percent and is considered to be a better measure of recognizer accuracy than the number of words correct. (I.e. words correct = 100% - deletions - substitutions).

2. Taking a transcript of TV broadcast data and re-recording it with an average reader in a speech lab under good acoustic conditions, using a close-talking microphone produces an estimated word error rate between 10 percent and 17 percent for speech recognition systems that were not tuned for the specific language and domain in question.

3. Speech, recorded by a professional narrator in a TV studio, which does not include any music or other noise gives us an error rate of around 20 percent. Part of the increased error rate is due to poor segmentation of utterances. There are places where the speech recognizer cannot tell where an utterance started or ended. This problem was not present in the lab recorded data. Different microphones and acoustics also contribute to the higher error rate.

4. Speech recognition on C-span broadcast data shows a doubling of the word error rate to 40 percent. While speakers are mostly constant and always close to the microphone, other noises and verbal interruptions degrade the accuracy of the recognition.

5. The dialog portions of broadcast documentary videos yielded recognition word error rates of 50 to 65 percent, depending on the video data. Many more environmental noises occur during outdoor recordings of speech.

6. The evening news was recognized with 55 percent error rate, which includes the extremely difficult to recognize commercials and introductions as well as the actual news program.

7. A full 1-hour documentary video including commercials and music raised the word error rate to 75 percent.

8. Worst of all were commercials, which were recognized with an 85 percent error rate due to the large amounts of music in the audio channel as well as the speech characteristics (and singing) in the spoken portion.

While these recognition results seem sobering at first glance, they merely represent a first attempt at quantifying speech recognition for broadcast video and audio material. Fortunately speech recognition does not have to be perfect to be useful in the Informedia digital video library. Informal retrieval experiments have indicated that redundancy in speech is a great help in retrieval. Speech errors also tend to be scattered throughout all documents, so that speech recognition errors will rarely result in spurious retrieval hits based on multiple query words.

Issues for Future Research

We are targeting four broad research areas: user interfaces, image understanding, natural language processing and speech recognition.

The user interface issues deal with the way users explore the library once it is available. Can the user intuitively navigate the space of features and options provided in the Informedia: News-on-Demand interface? What\ should the system provide to allow users to obtain the information they are looking for? Our plan is to move to a testbed deployment and gain insights from users as well as explore various interface design alternatives.

Natural language processing research for News-on-Demand has to provide acceptable segmentation of the news broadcasts into stories. We also want to generate more meaningful short summaries of the news stories in normal English. Natural language processing also has a role in query matching for optimal retrieval from the story texts. Finally, the system would greatly improve if queries could be parsed to separate out dates, major concepts and types of news sources.

Image processing research (Hauptmann & Smith 1995) is continuing to refine the scene segmentation component. The choice of a single key frame to best represent a whole scene is a subject of active research. In the longer term, we plan to add text detection and apply OCR capabilities to reading text off the screen. We also hope to include similarity-based image matching in the retrieval features available to a user.

Speech recognition helps create a time-aligned transcript of the spoken words. Speech recognition is inherently error prone and the magnitude and number of the errors determines whether the system is usable or useless. Thus recognizer accuracy is the critical factor in any attempt to use speech recognition for the digital video library.

When the speech recognizer does not find a trigram in the language model for the current hypothesized word triplet, it uses bigrams (word pairs) in the language model, although with a probability penalty. Similarly when an appropriate bigram cannot be found in the language model, individual word probabilities are used, again with a penalty. There were between 1 percent and 4 percent of the spoken words missing from the data. Since each missed word gives rise on average to 1.5 to 2 word errors, this alone accounts for 2 to 8 percent of the error rate. The bigrams in the language model were also inadequate. Depending on the data set, anywhere from 8 percent to 15 percent of the word pairs were not present in our language model. The trigram coverage gap was quite large. Between 70 percent and 80 percent of the trigrams in the data were not in the language model, so they would immediately be assigned with a much lower recognition probability.

By itself, the videos' unconstrained vocabulary degrades recognition. However, several innovative techniques can be exploited to reduce errors. The use of program-specific information, such as topic-based lexicons and interest-ranked word lists can be employed by the recognizer. Word hypotheses can be improved by using adaptive, "long-distance" language models and we can use a multi-pass recognition approach that considers multi-sentence contexts. Recent research in long distance language models indicates twenty to thirty percent improvement in accuracy may be realized by dynamically adapting the vocabulary based on words that have recently been observed in prior utterances.

In addition, most broadcast video programs have significant descriptive text available. These include early descriptions of the program, treatments, working scripts, abstracts describing the program, and captions. Closely related daily news text can be obtained from other sources of news such as the on-line wire services and newspapers. In combination, these resources provide valuable additional data for building recognition language models.

Speech recognizers are very sensitive to different microphones and different environmental conditions in which their acoustic models were trained. Even microphone placement is a factor in recognizer accuracy. The degradation of speech accuracy in the results of Table 1 between the lab and the broadcast data (using identical words) can be attributed to microphone differences and environment noise. We are actively looking at noise compensation techniques to ameliorate this problem. The use of stereo data from the left and right broadcast channel may also help in reducing the drop in accuracy due to environmental noise.

Perhaps the greatest benefit of speech recognition comes from alignment to existing transcripts or closed-captioning text. The speech recognizer is run independently and the result is matched against the transcript. Even though the recognition accuracy may only provide one correct word in five, this is sufficient to allow the system to find the boundaries of the story paragraphs.

In general, we distinguish 5 types of errors within News-on-Demand, all of which are subjects of active research:

1. False story segmentation happens when we incorrectly identify the beginning and end of a video paragraph associated with a single news story. Incorrect segmentation is usually due to inaccurate transcription, either because the closed-captioning itself has errors, because processing the closed-captioning text into stories is incorrect, or because of errors in the segmentation based on speech transcripts.

2. Incorrect words in the transcripts are either the result of faulty speech recognition or errors in the closed-captioned text. The result is the appearance of incorrect words in stories and consequent errors in the index and in retrieval.

3. False synchronization designates retrieved words that were actually spoken elsewhere in the video. This is due to closed captioning mismatched with the speech recognition.

4. Incorrectly recognized query. This is the result of an incorrect speech recognition during the library exploration. The user can edit and correct query misrecognitions through typing, or simply repeat or rephrase the query.

5. Incorrect set of stories returned for a query. This type of error is measured through information precision and recall. The user might get stories that are not relevant to the query or miss relevant stories. Some of these errors are the result of shortcomings in the database search engine.

Conclusions

Despite the drawbacks of a fully automated system, the benefits of News-on-Demand are dramatic. With News-on-Demand, we can navigate the complex information space of news stories without the linear access constraint that normally makes this process so time consuming. Thus Informedia: News-on-Demand provides a new dimension in information access to video and audio material.

Speech recognition will never be a panacea for video libraries. However, even speech recognition with reasonable accuracy can provide great leverage to make data accessible that would otherwise be completely unavailable. Especially in conjunction with the use of transcripts or closed-captioning, speech recognition even at high error rates is tremendously useful in the digital video library creation process. For queries, the ability to quickly correct and edit spoken commands makes the spoken query interface quite usable. The impact of News-on-Demand will be to broaden the ability to access and reuse of all standard news materials (e.g., TV, radio, text) previously generated for public broadcast.

References

Informedia. 1997 http://www.informedia.cs.cmu.edu/

Christel, M., Stevens, S., & Wactlar, H., 1994, "Informedia Digital Video Library," *Proceedings of the Second ACM International Conference on Multimedia*, Video Program. New York: ACM, October, pp. 480-481.

Stevens, S., Christel, M. and Wactlar, H. 1994, "Informedia: Improving Access to Digital Video". *Interactions* **1** (October), pp. 67-71

Christel, M., Kanade, T., Mauldin, M., Reddy, R., Sirbu, M., Stevens, S., and Wactlar, H., 1994, "Informedia Digital Video Library", Communications of the ACM", 38 (4), April, pp. 57-58.

Hauptmann, A.G., Witbrock, M.J., Rudnicky, A.I. and Reed, S., 1995, *Speech for Multimedia Information Retrieval*, UIST-95, Proceedings of User Interface Software Technology.

Hauptmann, A.G., Witbrock, M.J., and Christel, M.G., 1995, News-on-Demand: An Application of Informedia Technology, D-LIB Magazine, September 1995, http://www.cnri.reston.va.us/home/dlib.html, cnri.dlib/september95-hauptmann

CMU-SPEECH. 1997 "http://www.speech.cs.cmu.edu/speech", 1996.

CNN 1996, Cable News Network/Intel CNN at Work - Live News on your Networked PC Product Information http://www.intel.com/comm-net/cnn_work/index.html.

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

Brown, M.G., Foote, J.T., Jones, G.J.F., Sparck Jones, K., and Young, S.J., 1995, "Automatic Content-based Retrieval of Broadcast News," *Proceedings of ACM Multimedia*. San Francisco: ACM, pp. 35 - 43.

Hauptmann, A. G., and Smith, M., 1995,"Text, Speech, and Vision for Video Segmentation: The Informedia Project," *AAAI Fall 1995 Symposium on Computational Models for Integrating Language and Vision*, Cambridge, MA: MIT, pp. 90-95.

Mauldin, M., 1991, *Information Retrieval by Text Skimming*, Ph.D. Thesis, Carnegie Mellon University. August 1989. Revised edition published as "Conceptual Information Retrieval: A Case Study in Adaptive Partial Parsing, Kluwer Press.

Salton, G. and McGill, M.J., 1983, *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, McGraw-Hill Computer Science Series.

Zhang, H., Low, C., and Smoliar, S., 1995, "Video parsing and indexing of compressed data," *Multimedia Tools and Applications* **1**(March), pp. 89-111.

Hwang, M., Rosenfeld, R., Thayer, E., Mosur, R., Chase, L., Weide, R., Huang, X., and Alleva, F., 1994, "Improving Speech Recognition Performance via Phone-Dependent VQ Codebooks and Adaptive Language Models in SPHINX-II." *ICASSP-94*, vol. I, pp. 549-552.

Rudnicky, A., 1995, "Language Modeling with Limited Domain Data," Proceeding of the 1995 ARPA Workshop on Spoken Language Technology.