Speech Recognition in the Informedia^{TM} Digital Video Library: Uses and Limitations

Alexander G. Hauptmann
School of Computer Science, Carnegie Mellon University, Pittsburgh PA, USA

Abstract

In principle, speech recognition technology can make any spoken data useful for library indexing and retrieval. This paper describes the Informedia Digital Video Library project and discusses how speech recognition is used for transcript creation from video, alignment with hand-generated transcripts, query interface and audio paragraph segmentation. The results show that speech recognition accuracy varies dramatically depending on the quality and type of data used. Our information retrieval experiments also show that reasonable recall and precision can be obtained with moderate speech recognition accuracy. Finally we discuss some active areas of speech research relevant to the digital video library problem.

1. Introduction

Vast digital libraries of information will soon become available on the nation's Information Superhighway as a result of emerging multimedia computing technologies. However, it is not enough to simply 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 effectively reuse them.

This paper describes the Informedia Digital Video Library System [Christel94a, Stevens94, Christel94b, Informedia95], and the News-on-Demand application [Hauptmann95]. We show how speech recognition fits into the digital video library process. We present results for speech recognition on actual data. Finally we discuss the speech research issues that were uncovered as areas for further study.

2. The Informedia Digital Video Library Project

The Informedia Digital Video Library Project at Carnegie Mellon University is creating a large digital library of text, images, videos and audio data available for full content retrieval. Through the integration of technologies from the fields of natural language understanding, image processing, speech recognition and video compression, the Informedia System allows a user to explore multi-media data in depth as well as in breadth. An overview of the structure of the Informedia system is shown in Figure 1.

The Informedia Library project is developing these new technologies and embedding them in a video library system primarily for use in education and training. The nation's schools and industry together spend between $400 and $600 billion per year on education and training, an activity that is 93% labor-intensive, with little change in teacher productivity ratios since the 1800s. The new digital video library technology will bring about a revolutionary improvement in the way education and training are delivered and received.

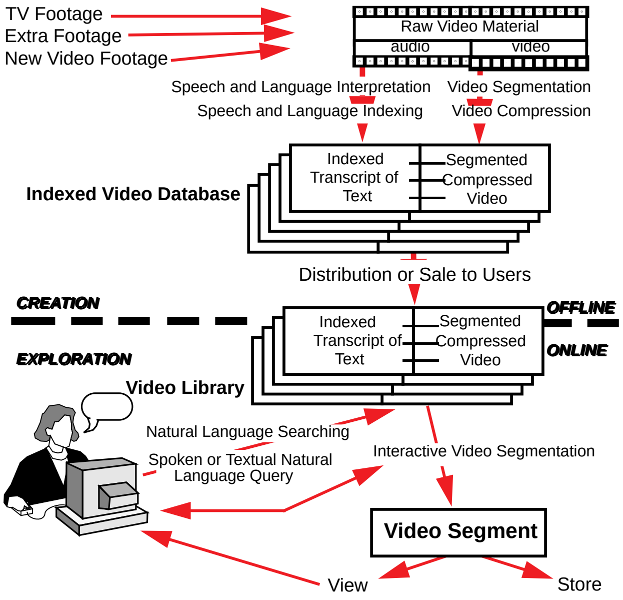
TV Footage
Extra Footage
New Video Footage
Raw Video Material
audio
video
Speech and Language Interpretation
Speech and Language Indexing
Video Segmentation
Video Compression
Indexed Video Database
Indexed Transcript of Text
Segmented Compressed Video
Distribution or Sale to Users
CREATION
OFFLINE
EXPLORATION
ONLINE
Indexed Transcript of Text
Segmented Compressed Video
Video Library
Natural Language Searching
Spoken or Textual Natural Language Query
Interactive Video Segmentation
Video Segment
View
Store

Figure 1. Overview of the Informedia Digital Video Library.

This material is based upon work supported by the National Science Foundation under Cooperative Agreement No. IRI-9411299. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

The Informedia system for video libraries goes far beyond the current paradigm of video-on-demand, retrieving and displaying short video paragraphs in response to the user's query. As a result, a large body of video material can be searched with very little effort.

We distinguish two distinct phases in the Informedia library process: library creation and library exploration.

Library creation deals with the accumulation of information, transcription, segmentation and indexing. Library exploration concerns the interaction with the user trying to retrieve selections in the database.

2.1 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 [HauptmannSmith95]:

a. *Text processing* looks at the textual (ASCII) representation of the words that were spoken, as well as other annotations derived from the transcript, the production notes or the close-captioning that may be available. Text analysis can work on an existing transcript to help segment the text into paragraphs [Mauldin89]. An analysis of keyword prominence allows us to identify important sections in the transcript. Other more sophisticated language based criteria are under investigation. We use two main techniques in text analysis.

1. If we have a complete time aligned transcript available from the close-captioning or through a human generated transcription, we can exploit natural "structural" text markers such as punctuation to identify segments of video paragraph granularity.
2. To identify and rank the contents of various segments, 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[Salton83].

b. *Image analysis* looks at the images in the video-only portion. Image analysis is primarily used for the identification of scene breaks and static frame icons that are representative of a scene. Image statistics methods compute primitive image features, such as color histograms, and use them for indexing, matching and segmenting images [Zhang95].

c. *Speech analysis* provides the basis for analyzing the audio component of the material, as discussed in the following sections.

In this paper we will focus on **speech analysis** for the digital video library.

2.2 Informedia Project Evaluation

Our project contains a plan for four test bed installations ranging from grade school children to university faculty. In addition, we will provide networked access to the primary test bed, and export portions of the system and data to other sites for their local exploration and experimentation.

The user tests will be established at Carnegie Mellon University, the Winchester Thurston School in Pittsburgh, the Fairfax County (VA.) public school system, and the Open University in the U.K. Users will be of many different types, as we test the viability of the library search concept and the usability of the user interface for various age and interest groups.

2.3 Informedia: News-on-Demand

A different application for the Informedia library system is News-on-Demand [Hauptmann95]. News-on-Demand monitors the evening news from the major networks and allows the user to retrieve stories of interest. 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. Immediate availability of the library information is important, as well as continuous updates of the contents.

Unlike other informedia prototypes which are designed to be educational test beds, the News-on-Demand system is fully automated. While the educational test bed prototypes develop their library with computer-assisted methods that require a certain amount of human post-processing, the News-on-Demand system is fully automatic. Thus we are forced to look at the potential for speech recognition without human corrections and editing.

3. Speech Recognition in Informedia

Speech recognition can aid the Informedia Digital Video Library system. both in library creation and during library exploration by the user.

3.1 Speech Recognition for Library Creation

Speech recognition during library (index) creation must consider data from the following sources:

- radio data,
- network broadcast video,
- video and telephone conferences, and
- movies and documentary videos.

In addition, transcripts or close-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 as well as segmenting 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. Accurate word boundaries are also necessary to create "video skims", which allow a user to skim through a video 5 to 20 times faster, but without missing important information.

Even though much of broadcast television is close-captioned, the vast majority of the nations video and film assets are not. More importantly, typical video production generates 50 to 100 times more data that is not broadcast and therefore not captioned. Clearly, effective use and reuse of video information assets will require automatic generation of transcripts.

Only a portion of the news stories have close-captioning text available. In any case, the close-captioned data, if available at all, lags up to 25 seconds behind the actual words spoken. This problem, as well as the problem of inaccuracies in transcription is especially glaring when the broadcast is “live”. In News-on-Demand we use speech recognition in conjunction with close-captioning to improve the time-alignment of the transcripts and to correct for gross errors.

For the news broadcasts that are not close-captioned, we need a transcript generated exclusively by the speech recognition system. The vocabulary and language model used here approximate a ``general American news” language model. It was based on a large corpus of North American business news from 1987 to 1994.

3.2 Speech Recognition for Library Exploration

Users are able to explore the Informedia library through an interface that allows them to search the Informedia library using typed or spoken natural language queries, select relevant documents retrieved from the library and play/display the material on their PC workstations. The library retrieval system can effectively process natural spoken queries and deliver relevant video data in a compact format, based on information embedded with the video during library creation. Video and other data may be explored in depth for related content. During retrieval based on keyword searches by a user, only the relevant video segments are displayed. During library exploration, the Sphinx-II speech recognition allows a user to query the system by voice, simplifying the interface by making the interaction more direct. The integration of speech with the interface enhances access to the stored video data by allowing more immediate and direct entry of queries.

The language model used during exploration is similar to the language model used during library creation, but augments it with the words actually found in the transcripts.

In the future we plan to limit the language model to only those words found in the actual library data, making the language model more efficient and smaller.

3.3 The Limitations of Speech Recognition

While it seems simple to hypothesize a speech recognizer for transcribing all the audio data into text, we must consider the accuracy of the recognition system. Speech recognition is inherently error prone and the magnitude of the errors will determine 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.

A second issue is the problem of segmentation. Speech recognizers are built to recognize individual sentences and can be modified to accept paragraphs. However, a full hour of continuous video data must be broken up into individual audio paragraphs through analysis of the speech signal. Ideally these audio paragraphs should contain complete sentences, or at least phrases. However, meaningful segmentation is still an open research issue.

Finally we must also consider the difference between speech recognition and speech understanding. Perfect speech recognition is not possible without understanding. Understanding natural language is an “AI-complete” problem, which means it is as difficult as creating artificial intelligence in general. For speech to be recognized perfectly, the recognizer needs to disambiguate what was said. This implies a close coupling to whatever natural language processing may be available for query processing or context determination.

4. The SPHINX-II Speech Recognition System

To transcribe the content of the video material, we use the Sphinx-II system, which is a large-vocabulary, speaker-independent, continuous speech recognizer created at Carnegie Mellon [CMU-Speech95, Hwang94]. 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, and 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 pos-

sible begin times for each end time predicted in the first pass.

3) An A* algorithm is used to generate 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 among the N-best list produced. This hypothesis is output as 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). The current 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 the most frequent 58800 words in the corpus [Rudnicky95].

5. Basic Speech Recognition Results

Table 1 shows the results from recognition experiments with different video data. The results on our data show that the type of data and the environment in which it was created dramatically alters the speech recognition accuracy.

Table 1: Speech Recognition Results

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

1) The basic reference point is the standard speech evaluation data which is used to benchmark speech recognition systems with large vocabularies between 5000 and 60000 words. The recognition 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 with an average reader and re-recording it in a speech lab under good acoustic conditions, with a close-talking microphone shows an estimate of 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 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. Here we have many more environmental noises as well as outdoor recordings of speech.

6) The evening news was recognized with 65 percent error rate.

7) A full 1-hour documentary video including commercials and music dropped 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 the usefulness of 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.

5.1 Speech Recognition Requirements for Information Retrieval

The transcript generated by Sphinx-II recognition need not be viewed by users, but can be hidden. However, the words in the transcript are time-aligned with the video for subsequent retrieval. Because of this, our system will tolerate higher error rates than those that would be required to produce a human-readable transcript.

We tested our assumption using 129000 queries on the data. The queries were collected from users looking for information from a World Wide Web index. Each query was applied to a perfect transcript of the documentary video data and to a SPHINX-II generated transcript of the same data. The results were scored in terms of information recall and precision. Recall was defined as the hits in the perfect transcript that were also returned by the search on the SPHINX-II generated transcript divided by the total hits in the perfect transcript. Precision was defined as the hits in the SPHINX-II transcript that were also in the perfect transcript, divided by the total number of hits in the SPHINX-II transcript.

.The results shown in Table 2 indicate that both recall and precision are very closely related to transcript word error rate. It is up to empirical tests to determine what levels of recall and precision are satisfactory for our users.

Table 2: Information Retrieval Results

Recognition Word Error Rate of Transcript	**Recall**	Precision
22 %	86 %	83 %
67 %	48 %	44 %
77 %	26 %	17.2 %

One consequence of errors is that incorrect stories are returned for a query. This type of error indicates lack of information retrieval recall and precision. The user might get stories that are not relevant to the query or miss relevant stories. Beyond that, incorrectly returned stories are the result of shortcomings in the processing of the query keywords.

Removing between 50 and 130 of the most frequent words of English did not improve recall or precision. This result indicates that it is not the case that most of the speech recognition errors are in common or function words.

5.2 The Speech Interface

Our current interface for the Informedia system is very simple. There are four buttons and a query text window. The text window shows the result of the speech recognition and can be edited by selecting portions of the text and typing.

The LISTEN button must be pressed while the user is speaking (push and hold while talking). Pushing the LISTEN button starts the recognition process and releasing the button signals the end of the query. In the future we hope to include a continuous listening mode to eliminate this button.

The SUBMIT button sends the text that is currently in the query text window to the search retrieval component, which then returns the relevant matches.

The CLEAR button erases all text from the text query window.

The UNDO button erases just the last recognized utterance from the query text window. All earlier utterances are still part of the current query until they are erased with a CLEAR.

This simple interface seems sufficiently simple and intuitive to learn. However we are currently experimenting with even fewer buttons, where UNDO becomes unnecessary (or a voice command) and every query is immediately submitted, eliminating the need for the separate SUBMIT button.

6. Speech Research Issues

There are a many speech issues that have been brought to light while studying this data and the uses of speech recognition in the Informedia Library. We will highlight only a few of these issues, that we consider especially relevant.

6.1 Language Models and Dictionaries

An analysis of the language models and dictionaries with respect to our data showed the following shortcomings. 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 word pairs (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 recognized with a much lower probability.

By itself, the video library's unlimited vocabulary degrades recognition rate. 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 additionally 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. Our 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 design called treatments, working scripts, abstracts describing the program, and captions. Words that are likely to appear in the daily news can be obtained from many other sources of news such as the on-line wire services. In combination, these resources can provide valuable additions to dictionaries used by the recognizer. We are exploring these ideas in our current research.

6.2 Microphones and Environmental Noise

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. Much of 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 and environment noise. We are actively looking a noise compensation techniques and microphone independence 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.

6.3 Alignment to Transcripts

Perhaps the greatest benefit of speech recognition comes from alignment to existing transcripts or close-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. It is also sufficient to allow the automatic creation of video skims, where the video fragments corresponding to the most important words are spliced together to produce a brief summary clip of the while video.

We also found that the close-captioning may itself be full of errors. Up to 50 % word error rate is evident when we compare the close-captions with a carefully hand-generated transcript. But the errors in the close-captioned text are either typographical errors or deliberate simplifications by the transcriber. Thus the errors are uncorrelated to the speech recognition errors and the combination of speech recognition and close-captioning can be effectively used.

6.4 Automatic Segmentation into Audio Paragraphs

We can detect transitions between speakers and topics which are usually marked by silence or low energy areas in the acoustic signal. To detect breaks between utterances we use a modification of Signal to Noise ratio (SNR) techniques which compute signal power. This algorithm computes the power of digitized speech samples where Si is a pre-emphasized sample of speech within a frame of 20

$$Power = \log\left(\left(\frac{1}{n}\right) \cdot \sum\left(Si^2\right)\right)$$

milliseconds. A low power level indicates that there is little active speech occurring in this frame (low energy). Segmentation breaks between utterances are set at the minimum power as averaged over a 1 second window. To prevent unusually long segments, we force the system to place at least one break within 30 seconds. This algorithm seems to be fairly robust in segmenting speech at silences or speaker changes. An empirical evaluation of the algorithm is in progress.

6.5 Music and Non-Speech Events

One area of research that is still quite new and promising concerns music and non-english speech events. This includes the detection of:

foreign languages,

non-speech sounds.

music, and

recognition of speech overlaid with noise/music.

At this point we have no results to report on these issues. However, progress in these areas would dramatically improve error rates.

7. Conclusions

There is no "listening typewriter." 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 close-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. Despite the drawbacks of errors in the system, the benefits of speech recognition are very dramatic.

In 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. In the future, we plan to add OCR capabilities for reading headlines and image processing for visual scene segmentation to the News-on-Demand system.

Universal access to vast, low-cost digital information and entertainment will significantly impact the conduct of business, professional, and personal activity. The initial impact of the project's activity will be on the broad accessibility and reuse of existing video materials (e.g., documentaries, news, vocational, training) previously generated for public broadcast; public and professional education; vocational, military and business training.

The greatest societal impact of what we do will most likely be in K-12 education. The Digital Video Library represents a critical step toward an educational future that we can hardly recognize today. Ready access to multimedia resources will bring to the paradigm of "books, blackboards, and classrooms" the energy, vitality, and intimacy of "entertainment" television and video games. The persistent and pervasive impact of such capabilities will revolutionize education as we've known it, making it as engaging and powerful as the television students have come to love.

Acknowledgments

I would like to thank Michael Witbrock for his help with News-on-demand and everything else, Mike Christel for the Informedia library interface, Ravi Mosur for the fbs6/fbs8 Sphinx-II implementation. I am also indebted to all the members of the Informedia Project under Howard Wactlar for helping to digitize and record the data as well as the CMU Speech Group under Raj Reddy for all their support.

8. References

[Christel94a] Christel, M., Stevens, S., & Wactlar, H. "Informedia Digital Video Library," *Proceedings of the Second ACM International Conference on Multimedia*, Video Program. New York: ACM, October, 1994, pp. 480-481.

[Christel94b] Christel, M., Kanade, T., Mauldin, M., Reddy, R., Sirbu, M., Stevens, S., and Wactlar, H.","Informedia Digital Video Library", Communications of the ACM", 38 (4), April 1994, pp. 57-58}

[Hauptmann95] Hauptmann, A.G., Witbrock, M.J., Rudnicky, A.I. and Reed, S., *Speech for Multimedia Information Retrieval*, UIST-95, Proceedings of User Interface Software Technology, 1995, in press.

[HauptmannSmith95] Hauptmann, A. G., and Smith, M., " Text, Speech, and Vision for Video Segmentation: The Informedia Project," *AAAI Fall 1995 Symposium on Computational Models for Integrating Language and Vision*, in press.

[Hwang94] Hwang, M., Rosenfeld, R., Thayer, E., Mosur, R., Chase, L., Weide, R., Huang, X., 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.

[Informedia95] http://fuzine.mt.cs.cmu.edu/im/informedia.html

[Mauldin89] Mauldin, M. "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, September 1991.

[Rudnicky95] Rudnicky, A., "Language Modeling with Limited Domain Data," Proceeding of the 1995 ARPA Workshop on Spoken Language Technology, in press.

[CMU-Speech95] "http://www.cs.cmu.edu/afs/cs.cmu.edu/user/air/WWW/SpeechGroup/Home.html"

[Salton83] Salton, G., McGill, M.J. "Introduction to Modern Information Retrieval," McGraw-Hill, New York, McGraw-Hill Computer Science Series, 1983.

[Stevens94] Stevens, S., Christel, M., Wactlar, H. "Informedia: Improving Access to Digital Video". *Interactions* **1** (October 1994), pp. 67-71

[Zhang95] Zhang, H., Low, C., and Smoliar, S. "Video parsing and indexing of compressed data," *Multimedia Tools and Applications* **1**(March 1995), pp. 89-111.