RJ 10143 (95018) April 22, 1999
Computer Science

Research Report

THE CUEVIDEO SPOKEN MEDIA RETRIEVAL SYSTEM

Savitha Srinivasan
Dragutin Petkovic
Dulce Ponceleon

IBM Research Division
Almaden Research Center
650 Harry Road
San Jose, California 95120-6099

Mahesh Viswanathan

IBM Research Division
T.J. Watson Research Center
Route 134
Yorktown Heights, NY 10598

IBM Research Division
Yorktown Heights, New York ▪ San Jose, California ▪ Zurich, Switzerland

RJ 10143 (95018) April 22, 1999
Computer Science

THE CUEVIDEO SPOKEN MEDIA RETRIEVAL SYSTEM

Savitha Srinivasan
Dragutin Petkovic
Dulce Ponceleon

IBM Research Division
Almaden Research Center
650 Harry Road
San Jose, California 95120-6099

Mahesh Viswanathan

IBM Research Division
T.J. Watson Research Center
Route 134
Yorktown Heights, NY 10598

ABSTRACT: The application of speech recognition technology has shown encouraging results for spoken media (audio/video) retrieval where the ave rage precision often approaches 80% of that achieved for perfect text transcriptions. It has also been shown that the performance of the retrieval system can be severely limited by the accuracy of the recognition system depending on the quality and content of the au dio. In this paper, we present the CueVideo spoken media retrieval system where the key contributions are in developing audio prefilt ering techniques to improve the accuracy of the transcript when applied to long unstructured media, and the development of an informa tion retrieval system designed to address specific characteristics of the transcript generated using speech recognition. We also pres ent experimental results which show that a combination of techniques we introduce can result in improved precision and recall over c urrent spoken media retrieval techniques for real world audio.

The CueVideo Spoken Media Retrieval System

Savitha Srinivasan, Dragutin Petkovic, Dulce Ponceleon
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120 USA
savitha, petkovic, dulce@almaden.ibm.com

Mahesh Viswanathan
IBM T.J. Watson Research Center
Route 134
Yorktown Heights, NY 10598 USA
maheshv@watson.ibm.com

ABSTRACT

The application of speech recognition technology has shown encouraging results for spoken media (audio/video) retrieval where the average precision often approaches 80% of that achieved for perfect text transcriptions. It has also been shown that the performance of the retrieval system can be severely limited by the accuracy of the recognition system depending on the quality and content of the audio. In this paper, we present the CueVideo spoken media retrieval system where the key contributions are in developing audio prefiltering techniques to improve the accuracy of the transcript when applied to long unstructured media, and the development of an information retrieval system designed to address specific characteristics of the transcript generated using speech recognition. We also present experimental results which show that a combination of techniques we introduce can result in improved precision and recall over current spoken media retrieval techniques for real world audio.

1 INTRODUCTION

Conventional information retrieval (IR) has mainly been focused on retrieving text documents from large collections of text documents. The basic principles of text retrieval are well known and have been evaluated thoroughly [Salton89]. Today, especially with the use of the Internet and advances in hardware and software, we are witnessing the proliferation of digital media (image, video, audio) and consequently, there is a greater need to allow easy and effective retrieval of such documents. There are several factors that make the indexing, search and retrieval of media documents challenging. They are:

- Media documents are usually very large digital files consisting of sampled noisy raw data (pixels, audio signals), not easily interpreted as symbols like ASCII characters. This makes even basic indexing operations very challenging.
- Such digital media data is often unstructured, i.e. information bearing parts of the signal (speech versus noise, objects versus background) are not easily detected.
- The number of media files is growing rapidly, making manual indexing less and less appealing.
- New IR, search and browse techniques are needed that take into account the large size, and nature of media and errors inevitably present in indexing [Aigrain96].

These problems greatly complicate both the indexing phase and the IR phase for media documents. Therefore, in many practical applications, we continue to rely on manual indexing of media with manual annotations, close captions etc. This makes the information retrieval problem for media prohibitively expensive and unreliable. We believe that the indexing phase of media is more complex and expensive than the IR phase and needs to be automated as much as possible.

In the CueVideo project at the IBM Almaden Research Center, we are developing a system to automatically analyze, summarize, and facilitate rapid retrieval and browsing of video and audio data. Our philosophy is to employ a set of automatic indexing techniques in order to make indexing and browsing large body of media practical. Since each method in isolation is not sufficiently reliable, we use a combination of such methods in an integrated manner to achieve desired retrieval accuracy. We use video indexing and IBM's ViaVoice(tm) speech recognition engine combined with audio filtering techniques for indexing the audio sections of media data. Our media data comprises of typical news clips, infomercials and instructional videos. The corresponding audio section can be considered as long (several minutes to an hour), unstructured and noisy (pure speech, music, combination of music, speech and noise). Applications of CueVideo are in all areas where manual indexing of videos and their full-length playback is not feasible, such as company training, distance learning, communications etc.

This paper focuses on the automated indexing and retrieval of audio information in the CueVideo system. Our goal is to automatically analyze and decode the audio content of the media, create an index of extracted words with timing information, and offer the user the ability to locate specific areas of interest in media data by constructing textual queries. The rest of this paper is organized as follows: Section 2 describes the related work in this area. Section 3 describes the different aspects of the CueVideo Spoken Media Retrieval System. Sections 4 and 5 describe the experimental evaluation and results.

2 RELATED WORK

We see the related work as falling into two categories: those related to methods for indexing, classification and retrieval of audio, and those related to the use of speech recognition for IR on spoken media documents.

2.1 Audio Analysis

The definition and extraction of the right set of features in order to index audio data is known to be a difficult problem. Searching for a particular sound or a class of sounds such as applause, violence, music or speech is one way of indexing an audio database [Wold96]. Another approach is to focus on one particular class of sounds, such as music or speech and apply indexing and retrieval techniques that are best suited for that class of sounds [Arons97, Pfeiffer96]. Real world audio from commercial TV, radio broadcasts or distance learning is seldom exclusively music or speech, however. It can at best be classified as predominantly being one or the other. In the CueVideo project, we use methods for audio analysis to segment the audio signal into sections of clear speech on which we use speech recognition to extract the transcript. The motivation for performing the audio analysis prior to speech recognition is to improve the accuracy of the transcript and the efficiency of the text IR component.

2.2 Speech Recognition

There are several ways in which speech recognition may be applied in order to enable the use of words as indexing and retrieval terms. If the query words are fixed and known prior to search time, the audio may be decoded against a fixed vocabulary consisting of the query words. This technique known as word spotting has the obvious shortcoming of being limited by the fixed number of query words. Large vocabulary continuous speech recognition (LVCSR) offers an alternative to this approach where the number of query words can be as large as the size of the vocabulary being used, typically around 65000. While this approach overcomes the limitations of word spotting, the vocabulary though large is still finite, and has a domain-specific language model [Schmandt94] component to it.

LVCSR systems typically consist of three components: a vocabulary, a language model and set of pronunciations for each word in the vocabulary. A language model is a domain-specific database of sequences of words in the vocabulary, along with the probabilities of the words occurring in a specific order. The language model assists the recognizer in decoding dictated speech by biasing the output of the speech system towards high probability word sequences. Recognizing out-of-vocabulary terms continues to be an open issue with this approach. The phone lattice scanning method [Robinson94] addresses recognition of out-of-vocabulary terms and can be used to find arbitrary terms consisting of any sequence of phonemes, though the accuracy may be lower than that of LVCSR systems.

Early efforts in spoken document IR based on a phone recognition system have resulted in 60-70% accuracy [Schauble95]. Word spotting [Jones95] techniques have been used successfully to perform retrieval for small, fixed keyword vocabulary yielding about 90% of the average text retrieval performance. The benefits of combining word and phoneme representation to improve IR were first evaluated by James [James96]. Spoken document retrieval by combining multiple index sources generated using large vocabulary speech recognition and phone lattice scanning have been reported to yield 85% of the retrieval performance of full text retrieval systems [Jones96]. The Informedia [Hauptmann95] project represents a pioneering effort in the use of speech recognition to generate a transcript for IR in the context of digital video libraries. They have shown that high speech recognition accuracy is not necessary to achieve IR effectiveness similar to that of perfect text. They also propose an improved IR technique for imperfect transcripts by combining a word document index with a phonetic transcription index [Witbrock97].

The successful word spotting systems are obviously constrained by the size of the fixed word vocabulary, and therefore do not yield a general purpose retrieval solution for spoken documents. The state-of-the-art automatic phone recognition technology is at best about 70% accurate [Robinson94]. Also, the phone lattice system works with exact word forms and not stems. The best retrieval performance so far appears to be with combined indexes using large vocabulary speech recognition and phoneme based methods. These methods are effective when the audio being indexed consists mainly of well recorded speech segments. Real world audio which includes dialogs, broadcast news and commercials poses a problem since the recognition word error rate is typically high, leading to poor IR performance [Hauptmann95].

2.3 Spoken Media Retrieval Challenges

Retrieval systems which are based on the output of speech recognition technology are faced with several issues. Real world audio seldom consists of well recorded speech only. The speech segments that do exist do not contain any explicit verbal representations for the punctuation. We therefore lose any information on sentence structure. There is also the issue of language model mismatch for the audio content being indexed.

In this paper, we wish to explore and evaluate the extent to which we can improve on the use of LVCSR for spoken media retrieval over existing methods to retrieve relevant audio (and therefore video) clips based on keyword search. We divide the problem of IR for spoken media documents into challenges associated with indexing and challenges associated with IR. In indexing, we focus on segmentation of the audio signal into pure speech segments where speech recognition can be successfully applied, and the adaptation of speech recognition for indexing of large unstructured audio documents. On the IR side, we focus on techniques that can compensate for typical errors that occur when indexing transcript generated using speech recognition. Our spoken media retrieval system uses IBM's large vocabulary (65000 word) continuous speech recognition system (ViaVoice(tm)) with the broadcast news language model [Polymenakos98] for different types of audio [Bahl94, Bahl95]. The system runs at real-time or better on a 266MHz Pentium II personal computer. The word error (deletions, insertions and substitutions) rates under different audio conditions (speech with background music, speech with background noise etc.) for untrained speakers vary between 22% and 30%.

3 THE CUEVIDEO SPOKEN MEDIA RETRIEVAL SYSTEM

The system consists of an off-line indexing phase followed by an on-line information retrieval phase. During the indexing phase, we extract the audio track from the video and perform the audio analysis to detect speech boundaries. LVCSR is applied to the entire audio track to generate a transcript, of which only the speech segments are indexed. We also analyze the transcript to automatically extract the information bearing words. The retrieval phase consists of a user interface to select, construct and process the query, and display the results as a ranked list with the ability to play the corresponding section of video/audio.

3.1 Analysis and Segmentation of Audio

Audio is traditionally described by its pitch, loudness and duration. These attributes correspond to measurable features in the audio signal such as amplitude, frequency and phase. Changes in these features are perceived by humans as changes in loudness and pitch. Since every sound is composed of different frequencies and amplitudes whose change pattern is unique, the duration of such patterns is a first step towards segmentation and classification of audio.

We use a combination of features in the audio such as the zero crossing rate (ZCR), spectral concentration and harmonics to identify the boundaries of *clear speech* segments and *non-speech* segments in the audio data. We analyze the audio signal in both the temporal domain and the frequency domain to compute these features. The frequency domain information or the spectrum is obtained by performing a Fast Fourier Transform on the audio signal. The ZCR is a measure of the number of times in a given time interval that the speech signal amplitude passes through a value of zero and is computed from the signal in the time domain. Speech signals produce a marked rise in ZCR occurring at the beginning and ending of words due to the presence of consonants, a high ZCR is therefore a good indicator of the presence of speech. Speech is usually limited in frequency to 8 kHz whereas the frequencies in music can extend upto 20 kHz. The computation of the dominant frequency of the audio signal based on the spectral analysis is another distinguishing factor between speech and music. Also, the presence of harmonics (a group of frequencies simultaneously present for a long time) may be used as an indicator of music. We use empirically determined thresholds for the ZCR, the dominant frequency and similarity measures for groups of frequencies to distinguish speech segments from non-speech segments.

One interesting aspect of this speech boundary detection is the estimation of the duration of the speech versus non-speech segments. Our segmentation algorithm is able to detect speech segments that are as short as 100 ms. However, it is the application that determines the length of a meaningful duration for a segment. For example, use of detected speech boundaries for indexing of speech segments requires a segmentation granularity of 8 seconds or higher since the speech recognizer outputs on an average, 60-100 words per minute. Other applications that rely on audio information such as genre classification, audio summarization and skimming may require a lower temporal granularity of segmentation. Table 1 summarizes the performance of our audio analysis algorithms where the accuracy is calculated as shown below. The ground truth was computed based on manual observation of the start and end times of the non-speech segments.

Detection Accuracy = $\frac{N_D}{N_T}$ where

N_D is the number of detected segments that match the ground truth with a tolerance of +/- 1 second

N_T is the total number of detected segments

$$\text{Segmentation Accuracy} = \frac{\sum \frac{(|S_k(d)-S_k(g)|)+(|E_k(d)-E_k(g)|)}{E_k(g)-S_k(g)}}{N}$$

where k in the summation varies between 1 and N

$S_k(d)$ is the start time of the detected segment

$S_k(g)$ is the start time of the corresponding segment in the ground truth

$E_k(d)$ is the end time of the detected segment

$E_k(g)$ is the end time of the corresponding segment in the ground truth

N is the number of detected segments

From table 1, the average detection accuracy is found to be 80% and the average segmentation accuracy to be 88%. In our test set, almost all the detection errors were due to false positives, which may be seen as achieving recall performance of 1 since each segment in the ground truth was also detected by our algorithm. The detected false positives and our definition of detection accuracy contribute to the detection accuracy being less than 100%. The segmentation accuracy ignores the detected segments that have no corresponding segments in the ground truth. For the detected segments that do have a corresponding segment in the ground truth, it provides a measure of how closely the detected segments match the corresponding segments in the ground truth.

Video	Detection Accuracy	Segmentation Accuracy
Marketing Video on IBM Patent Server	75%	71%
Instructional video on coffee machine	100%	93%
Marketing video on IBM's radiology dictation product	67%	97%
News clips on IBM's Deep Blue Computer	85%	94%

Table 1: Audio Analysis Accuracy

3.2 LVCSR for Automated Indexing

The ViaVoice speech recognition engine is used to transcribe the audio and generate a continuous stream of words. For indexing, we define a unit-document to be a 100 word segment where consecutive segments overlap partially in order to address the boundary conditions. There are several operations performed in sequence in this processing. First, the words and times from the recognizer output are extracted to create the unit-document files with associated timestamps. The Julian time at the start of the audio is used as the reference basis. The time information assists in developing a user interface which not only displays the document relevant to the query, but can be extended to play the corresponding clip from the video or audio file. This is followed by tokenization to detect sentence/phrase boundaries and then part-of-speech tagging such as noun phrase, plural noun etc. The morphological analysis uses the part-of-speech tag and a morph dictionary to reduce each word to its morph. For example, the verbs, *lands*, *landing* and *land* will all be reduced to *land*. Then, the stop words are removed using a standard stop-word list. For each non-stop word, the number of unit-documents that it belongs to (the inverse document frequency) is computed and is used to weight the non-stop word.

3.3 Retrieval

The retrieval system first loads the inverted index and the precomputed weights of each of the non-stop words. A single pass approach [Robertson95] is used to compute a relevancy score with which each document is ranked against a query. The relevancy score is given by the Okapi formula:

$$S(d,q) = \sum \frac{(C_q(q_k)*C_d(q_k)*idf(q_k))}{a_1+(a_2*(\frac{l_d}{l_{bar}}))+C_d(q_k)}$$

where k in summation varies between 1 and Q

$S(d,q)$ is the score of document d against query q

q_k is the *kth* term in the query

Q is the total number of terms in the query

$C_q(q_k)$ is the count of the *kth* term in the query

$C_d(q_k)$ is the count of the *kth* term in the document

$l_d$is the length of the document

l_{bar} is the mean length of the documents in the collection

$a_1 = 0.5$

$a_2 = 1.5$

$idf(q_k)$ is the inverse document frequency for the query term q_k where

$$idf(q_k) = \log(\frac{(N-n(q_k)+0.5)}{(n(q_k)+0.5)})$$

where

N is the total number of documents

$n(q_k)$ is the number of documents that contain the term q_k

Each word in the query string is tokenized, tagged, morphed and then scored using the Okapi formula above. The total relevancy score for the query string is the combined score of each of the query words. The scoring function takes into account the number of times each query term occurs in the document normalized with respect to the length of the document. This normalization removes bias that generally favor longer documents since longer documents are more likely to have more instances of any given word. This function also favors terms that are specific to a document and rare across other documents. We minimize insertion and substitution word errors by filtering out the results that contain only short query words (<4 characters long) since short words are more likely to have been recognized in error as compared to longer words with distinct acoustics.

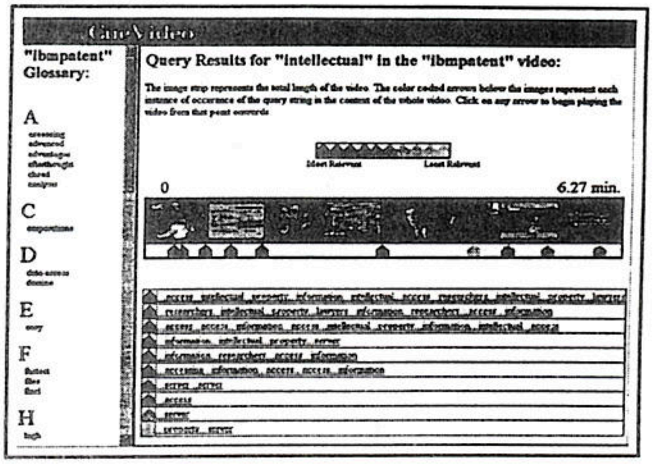
3.4 Extraction of Information Bearing Words

We automatically extract the common information bearing words in the transcript and use them as indicators of the main theme of the document in the user interface. To do this, we are restricted to frequency keyword heuristics since we have no information on sentence boundaries and document structure, and therefore the *pattern* of occurrence of the word in the document. We therefore adopt a practical filter solely based on frequency of occurrence as suggested by Luhn [Luhn57].

We use the following method to extract the words: We query the retrieval system using each word in the generated transcript. For each query word, the retrieval system returns a list of relevant documents based on the calculated Okapi score which takes the frequency of the query word into account. We prune the query results based on empirically determined thresholds for the document frequency and the query word frequency within each document. For example, we attempt to eliminate stray speech recognition word errors by thresholding the minimum number of document hits to be above a certain value. We also refine the query by eliminating very short query words.

3.5 User Interface

The web-based retrieval interface in the CueVideo system consists of several components. The extracted information bearing words are displayed in the query window to construct the query and serve to help the user both in constructing the query and obtaining some kind of textual summary of the document. The query is constructed by selecting the video and typing in the query word/phrase. Figure 1 shows a screen shot of retrieved results after a query has been processed. The left column in figure 1 continues to display a alphabetical listing of the information bearing words. These words may also be displayed in decreasing order of frequency of occurrence or on a temporal scale. The query results are visually presented as a movie summary with color coded arrows pointing to the places where the corresponding queried words or phrase occurs. Color coding of arrows serves to denote the relevance of the search result. The media (video/audio) can be played starting from the approximate time of the detected words.



Query Results for "intellectual" in the "ibmpatent" video:

The image strip represents the total length of the video. The color coded arrows below the images represent each instance of occurrence of the query string in the context of the whole video. Click on any arrow to begin playing the video from that point onwards

Figure 1: User Interface to Present Query Results

4 SPOKEN MEDIA RETRIEVAL EXPERIMENTS

We conducted a series of retrieval experiments in order to evaluate the impact of our enhancements to the retrieval system. We formulated a set of queries spanning across four different videos adding up to about 40 minutes. The statistics for our test data is given in table 3.

4.1 Evaluation Measures

We refer to the base spoken media retrieval system as the one which does not include the audio analysis for speech boundary segmentation. Our objective is to evaluate the effect of our improvements to the base system, it is *not* to evaluate the performance of the base system in comparison with an equivalent full text retrieval system. There are several interesting aspects which must be kept in mind while evaluating this system. First, we index each video/audio clip separately, hence our indexes do not span across all the video/audio clips to create a single, combined index. Since the clips vary from one another in language model match, quality of audio and percentage of speech versus non-speech segments, our current evaluation maintains separate indices. The maximum number of documents (100 word segments) in any given video from our test data set is fairly small (43). We have introduced the concept of *temporal proximity* in the retrieval system such that a query result which is temporally too close (empirically determined to be around one minute) to a previous query result is eliminated from the result set. The justification for this is that if a user is viewing a previous query result and the content is relevant/interesting, the user will continue to view the video at that point. This results in our evaluating the retrieval system which is not necessarily outputting the top N hits. The ground truth for the queries was determined by user observation of the video/audio where one or more occurrences of a query word within a 30

Video	Query Words	Length in minutes	Number of 100 word documents	Relevancy to Broadcast News Language Model	Number of "non-speech" segments > 8 seconds
Marketing Video on IBM Patent Server	intellectual, portfolio, server, lotus	6.27	15	Low	4
Instructional video on coffee machine	steam, milk, cup, grounds, bean	9.48	32	Very Low	3
Marketing video on IBM's radiology dictation product	dictation, speak, reports, radiology	9.45	30	Medium	2
News clips on IBM's Deep Blue Computer	champion, deep blue, Kasparov, human, research	12.63	43	High	5

Table 2: Statistics for Test Data Used in Retrieval Experiments

second window was considered to be a single relevant segment. The recall and precision are calculated as follows:

Recall = $\frac{N_{RR}}{N_T}$ where

N_{RR} is the number of relevant segments retrieved

N_T is the total number of relevant segments in the video based on ground truth

Precision = $\frac{N_{RR}}{N_R}$ where

N_R is the number of segments retrieved

4.2 Audio Analysis Filter

We have experimented with indexing and retrieval of audio segments that consist of several music/noise segments mixed with speech segments. We observed mixed results for our test data and believe that the benefits of this technique will be more significant for audio content that contains longer non-speech segments. The longest non-speech segment in our test data set was about 10 seconds. The transcript corresponding to a 10 second audio segment consists of about 10 words. The specific characteristics of these words when combined with the short duration of the segments does not result in a conclusive impact on the retrieval performance. Our specific observations are summarized in section 5 below.

4.3 Query-Expansion Techniques

In order to improve the recall of the retrieval system, we have explored means of expanding the query. To this end, we use the Textract [Vaithyanathan98] technology where multi-word terms are identified in text by a cascaded sequence of extractors. These extractors produce a set of terms for a collection of documents by scanning the document text and identifying linguistic expressions that refer to important concepts or entities mentioned in the text. Since the terms are derived from the texts, it is guaranteed not to offer terms not appearing in the texts. Furthermore, with effective text analysis techniques, it offers many of the terms which do occur and which are therefore relevant for applications such as document clustering that exploit the specific content of the corpus. We use this technology with a limited objective of automatically extracting phrases or co-occurring terms from the transcript. We expand the original query to include all the phrases that consist of the original query term, often resulting in an increased frequency of the individual query terms in the query phrase. This has a direct positive impact on the retrieval performance since the relevance scoring formula has a dependency on the frequency of the query terms.

We have also investigated the use of a phonetic thesaurus for query-expansion. However, we run into the following problem: Continuous speech recognition systems allow the user to pace their speech naturally. This causes adjacent words to run into each other, changing the acoustics within each word, based on the preceding and subsequent spoken words. This is referred to as coarticulation. As a result, each word does not have a "clean", or isolated, pronunciation associated with it, hence the difficulty in identifying phonetic errors. Therefore, phonetic errors, while appearing to be promising measures for query-expansion are not practical.

Table 3 lists the detailed results of the experiments conducted on the retrieval system.

5 LESSONS LEARNED

From table 3, we observe that the average recall performance of the base system is 0.51. This is directly influenced by the speech recognition accuracy as validated by our ground truth observation where the generated transcript was also visually scanned for the query words.

Video	Base System		Base System with Audio Analysis Filter		Base System with Expanded Queries		Base System with Audio Analysis Filter and Expanded Queries	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
IBM Patent Server	1	0.72	1	0.42	0.76	0.97	0.76	0.97
Coffee Machine	1	0.43	1	0.43	0.85	0.83	0.85	0.83
Radiology Dictation	0.93	0.29	1	0.26	0.95	0.64	0.95	0.64
IBM's Deep Blue	0.85	0.61	0.94	0.68	0.75	1	0.84	1

Table 3: Detailed Retrieval Results

The low recall was due to the query words not being present in the transcript as opposed to being present and not being retrieved by the retrieval system. However, the average precision of the base system is quite good.

Based on our initial experiments, we make the following observations:

- For our test data, the audio analysis filter resulted in marginal improvement in precision, where we had expected a higher improvement in precision. Further investigation of this issue reveals the following reason for this: The text in the transcript that corresponds to non-speech segments in audio mainly consists of *repeated, short, non information bearing words* that are unlikely to be valid index or query terms (such as "and", "half" etc.). It consists of very few information bearing words that are valid index terms, and the ones that were found were *unlikely* query terms for the media being indexed. For example, "ironic", "hour-and-a-half" and "Dallas" were some of the words in the transcript corresponding to the non-speech segments. Therefore, we conclude that the audio analysis filter is more appropriate for audio data that consists of long (>10 Sec), mixed segments of speech and non-speech. For such data, it improves the efficiency of indexing. In addition, the audio analysis filter can significantly improve the precision for content-based retrieval systems by providing cues to determine valid speech segments that are not to be indexed. For example, the elimination of a two minute commercial segment that contains a mixture of speech and music following a ten minute speech-only broadcast news segment will improve the retrieval precision.
- Our query-expansion technique significantly improves recall without severe degradation in precision performance. This indicates that efforts to correct the transcript or generate additional index terms for the transcript that attempt to compensate for the speech recognition errors will certainly improve recall. At this time the query is expanded strictly based on the text transcript. Using statistical information from the speech recognition engine such as confidence levels and alternate words to assist query-expansion is a promising direction to pursue.
- The combination of the audio filter with the query-expansion results in good performance numbers for precision and recall.
- Language model match for speech recognition does not appear to be a dominant factor in retrieval system performance. In our test data set, the content of News clips on IBM's Deep Blue Computer listed in table 1 is the best match to the broadcast news language model. However, the results tabulated in table 3 do not show a marked poor performance for the other videos as compared to the one with the best content match.
- Most speech recognition errors appear to be word deletions/substitutions rather than content bearing word insertion errors that may end up as incorrect index terms. This results in lowering the recall of the retrieval system rather than the precision. Therefore, a focus on addressing the word deletion/substitution errors using query-expansion techniques is likely to improve the performance of the IR system. This is consistent with Informedia's [Hauptmann95] findings where the removal of common words from the transcript did not improve precision or recall.

6 CONCLUSION AND FUTURE WORK

We have described our experience with using large vocabulary speech recognition for spoken media retrieval. We believe we achieved encouraging results with respect to our primary goal of attempting to compensate for imperfect recognition technology in the analysis and retrieval components. Elimination of short query words, performing audio analysis for segmenting speech from non-speech sections, indexing and query-expansion techniques coupled with good user interfaces were effective in improving the performance of our spoken media retrieval system.

As we continue this work, we would like to introduce a statistical component to the text information retrieval component based on an understanding of the nature of speech recognition errors. Two specific statistical measures that speech recognition systems can provide are confidence levels associated with the recognition of a word and lists of next best guesses for each recognized word. This information can be used for weighting the index terms, query-expansion, retrieval and creating additional index terms. We would also like to explore means of addressing the out-of-vocabulary retrieval problem in LVCSR systems using query-expansion/refinement techniques. Another aspect of the future work involves analysis of the speech segments to explore patterns in the frequency or temporal domain to make reasonable interpretations at a level higher than the word level. A combination of detected hesitations, pauses and emphasis in the speech may be used to influence the weight of the index terms such that words recognized in an emphasized speech segment are weighted higher. Finally, we would like to validate our retrieval performance with additional test data and more detailed evaluation methods.

7 ACKNOWLEDGMENTS

We thank Byron Dom for his advice on effective evaluating measures for spoken media retrieval systems and thank Brian Blanchard for conducting the retrieval experiments. We also thank Norm Pass for his encouragement and support.

REFERENCES

[Aigrain96] Aigrain, Zhang and Petkovic. Content-Based Representation and Retrieval of Visual Media: A State-of-the-Art Review. In *Multimedia Tools and Applications, Vol. 3, pp. 179-202,* Kluwer Academic Publishers, 1996.

[Arons97] Arons, B., SpeechSkimmer: A System for Interactively Skimming Recorded Speech. In ACM Transactions on Computer-Human Interaction, *Volume 4, Number 1, pp. 3-38*.

[Bahl94] Bahl, L.,R. et al., Robust Methods for using Context-Dependent features and models in a continuous speech recognizer. In Proceedings of *ICASSP 94*.

[Bahl95] L. R. Bahl et al., Performance of the IBM large vocabulary continuous speech recognition system on the ARPA Wall Street Journal task. In *Proceedings of ICASSP 95, pp 41-44*.

[Hauptmann95] Hauptmann, A.G. Speech Recognition in the Informedia Digital Video Library: Uses and Limitations. In *Proceedings of ICTAI-95 7th IEEE International Conference on Tools with AI*, Washington, DC.

[Jones95] Jones, G. J. F., Foote, J. T., Jones, K. S., and Young, S. J.. Video Mail Retrieval: the effect of word spotting accuracy on precision. In Proceedings of *ICASSP 95, volume 1, pp. 309-312*, Detroit, MI.

[Jones96] Jones, G. J. F., Foote, J. T., Jones, K. S., and Young, S. J. Retrieving Spoken Documents by Combining Multiple Index Sources. In *Proceedings of SIGIR 96, pp. 30-38*, Zurich, Switzerland.

[Luhn57] Luhn, H.P., A Statistical Approach to the Mechanized Encoding and Searching of Literary Information. In *IBM Journal of Research and Development 1(4), pp. 309-317*.

[Pfeiffer96] Pfeiffer, S., Fischer, S. and Effelsberg, W. Automatic Audio Content Analysis, in *Proceedings of MM'96, pp. 21,* Acm Press.

[Polymenakos98]L. Polymenakos, P. Olsen, D. Kanewsky, R.A. Gopinath, P.S. Gopalakrishnan, and S. Chen. ' Transcription of Broadcast News - Some Recent Improvements To IBM's LVCSR System. In Proc. *ICASSP 98,* Seattle, WA.

[Robertson95] Robertson, S.E., Walker, A., Sparck-Jones, K., Hancock-Beaulieu M.M & Gatford, M. Okapi at TREC-3. In Proc. Third Text Retrieval Conference. (NIST special publication), 1995.

[Robinson94] Robinson, T., Hochberg, M., and Renals, S. IPA: Improved phone modelling with recurrent neural networks. In Proc. *ICASSP 94, volume 1, pp. 37-40,* Adelaide, SA.

[Salton89] Salton, G. Automatic Text Processing, Reading, MA, Addison-Wesley, 1989.

[Schauble95] Schauble, P. and Wechsler, M. First Experiences with a System for Content Based Retrieval of Information from Speech Recordings. In *IJCAI-95, Workshop on Intelligent Multimedia Information Retrieval*, Maybury, M.T.

[Schmandt94] Schmandt, C. Voice Communication with Computers. Van Nostrand Reinhold, New York 1994.

[Vaithyanathan98] Vaithyanathan, S., Ravin, Y., Byrd, R. and Dhillon, I. Automatic Labeling of Document Clusters. Upcoming IBM Technical Report.

[Witbrock97] Witbrock, M. and Hauptmann, A. Using Words and Phonetic Strings for Efficient Information Retrieval from Imperfectly Transcribed Spoken Documents. In *Proceedings of DL97, The Second ACM International Conference on Digital Libraries,* Philadelphia, PA.

[Wold96] Wold, E., Blum, T., Keislar, D. and Wheaton, J.. Content-Based Classification, Search, and Retrieval of Audio. In *IEEE Multimedia*, *volume 3, No. 3, pp. 27-36.*

ViaVoice is a registered trademark belonging to the IBM Corporation