

# SpeechBot: a Speech Recognition based Audio Indexing System for the Web

**Jean-Manuel Van Thong, David Goddeau, Anna Litvinova<sup>1</sup>,  
Beth Logan, Pedro Moreno, Michael Swain**

Cambridge Research Laboratory, Compaq Computer Corporation  
One Cambridge Center, Cambridge, MA 02142, USA  
{jmvt,btl,pjm}@crl.dec.com

## Abstract

We have developed an audio search engine incorporating speech recognition technology. This allows indexing of spoken documents from the World Wide Web when no transcription is available. This site indexes several talk and news radio shows covering a wide range of topics and speaking styles from a selection of public Web sites with multimedia archives. Our Web site is similar in spirit to normal Web search sites; it contains an index, not the actual multimedia content. The audio from these shows suffers in acoustic quality due to bandwidth limitations, coding, compression, and poor acoustic conditions. The shows are typically sampled at 8 kHz and transmitted, RealAudio compressed, at 6.5 kbps. Our word-error rate results using appropriately trained acoustic models show remarkable resilience to the high compression, though many factors combine to increase the average word-error rates over standard broadcast news benchmarks. We show that, even if the transcription is inaccurate, we can still achieve good retrieval performance for typical user queries (69%). Because the archive is large - over 5000 hours of content (and growing at a rate of 100 hours per week), totaling 47 million words and growing rapidly - we measure performance in terms of the precision of the top-ranked matches returned to the user.

## 1. Introduction

The amount of audio and video content available on the Internet has increased dramatically over the past few years. Unfortunately, indexing engines are often limited to text and image indexing and many multimedia documents, video and audio, are thus excluded from classical retrieval systems. Very often, there is no transcription available for these documents and therefore no simple means for indexing their content. Except for some carefully archived shows at the National Public Radio site (NPR), most archives of radio shows on the Web are simple lists of links to long audio files, sometimes several hours in length (Broadcast, 1999). A searchable index that provides the ability to play the segments of interest within the audio file of these shows would make these archives much more accessible for listeners interested in a particular topic.

A straightforward approach to solve this problem consists of generating the transcription automatically using a large vocabulary speech recognition system. However, speech recognition technology is currently inherently inaccurate, particularly when the audio quality is degraded due to poor recording conditions and compression schemes. Despite this, we show that we can achieve accuracy satisfactory for indexing audio from the Web if the acoustic and language models are properly trained. There have been several studies which had similar goals (Wactlar, 1996; Garfalo *et al.*, 1998; Johnson *et al.*, 1999; Abberley *et al.*, 1999). We differ from these projects in several ways. First, we fetch the audio documents from the Web and build an index from that data. Second, we don't serve content, but rather keep a link to the original document, similar to traditional search engines. Finally, our system is designed to scale up on demand.

In many studies, document retrieval has been shown to be remarkably resilient to word error rate. A recent study (Witbrock, 1999) shows that a word error rate of 30% reduces recall by 4%, and a word

---

<sup>1</sup> Anna Litvinova is a Harvard University student. This work was performed while completing an internship at CRL.

error rate of 50% reduces it by only 10% (see also Garfalo *et al.*, 1998). There are several explanations for these numbers. When the recognizer misses a word once, it may still be recognized other times if it appears in the same audio document. Also, if there are many words in the query, missing one or two of them may still permit retrieval of the document.

The effectiveness of retrieval is affected by a number of factors such as the complexity of the queries, the length of the documents, the size of the collection, and the information retrieval techniques used. Because Web searchers are known to rarely look past the first page or two of results from a search engine, we are most concerned about the precision of the top matches (Silverstein *et al.*, 1998).

The content currently indexed is popular talk radio, technical and financial news shows. These shows are almost entirely speech, and they do not have associated transcriptions, unlike TV shows that are often closed captioned in the US. Closed captions (or CC) are the textual transcriptions of the audio track of a TV program, and they are similar to subtitles for a movie show. The Federal Communications Commission (FCC) requires broadcast programs to be close captioned in order to be accessible to people with disabilities. These closed captions can easily be used for indexing purposes since the transcription is time aligned for display purpose.

Some of the indexed shows are extremely popular (Broadcast, 1999): Dr. Laura is syndicated to 430 stations and is reported to have a weekly audience of 18 million people; Art Bell's Coast to Coast/Dreamland is reported to have a weekly audience of 10 million people. Each Art Bell's Coast-to-Coast show is four hours long, and each Dr. Laura show is two hours long. However, none of these shows have existing indexes, except in some cases a title for each show. Long shows may cover a variety of different subjects and should not be indexed as a whole. Since these shows all have existing archives on the Internet, SpeechBot provides links to the content in these archives.

This paper describes first the overall architecture of the system. We then present a performance analysis. Finally we conclude with suggestions for improvements and directions for future research.

## **2. System overview**

SpeechBot is a public search system (SpeechBot, 1999) similar to AltaVista which indexes audio from public Web sites such as Broadcast.com, NPR.org, Pseudo.com, and InternetNews.com. The index is updated daily as new shows are archived in their Web sites. The system consists of the following modules: the transcoders, the speech decoders, the librarian database, and the indexer as shown in Figure 1. We describe each module in the following paragraphs.

### **Transcoder**

The transcoders fetch and decode video and audio files from the Internet. For each item, it extracts the meta-data, downloads the media documents to a local repository and converts the audio into a common uncompressed file format (typically Microsoft PCM wav sampled at 8 kHz). The meta-data contains information about the file downloaded such as the sample rate, copyright, the story title, and possibly a short description. This information is used by the Librarian database to identify and track the document while processed by the system, and to display the result of the query to the user.

The current version of the transcoders can handle a variety of different formats. Most of the documents downloaded however are RealAudio encoded. Transcoders run in parallel on a farm of 8 Compaq AP400 dual 400 MHz Pentium II workstations, 256 Mb RAM, under Windows NT. Each machine can handle 4 streams in real-time, leading to a maximum throughput of 768 hours of audio per day.

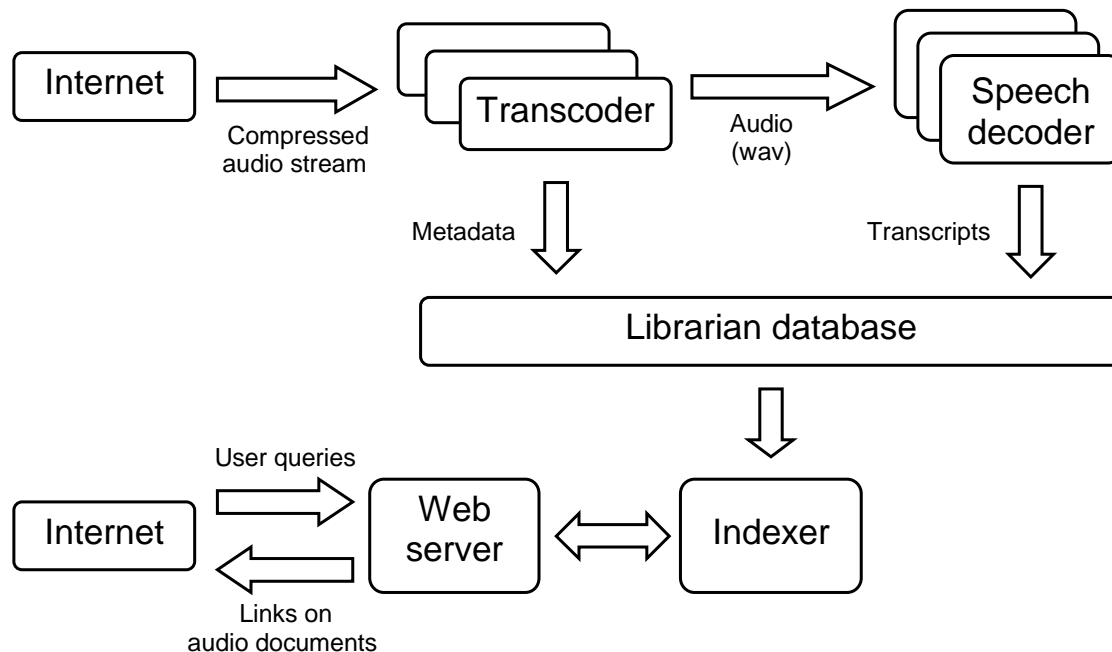


Figure 1: Overall architecture of the system.

### Speech Decoder

The Calista system is a large vocabulary continuous speech recognition package using state-of-the-art Hidden Markov Model (HMM) technology developed at the Compaq Cambridge Research Laboratory (CRL). It produces a textual transcription, or annotation stream, from the downloaded audio files. The annotation stream consists of the start time and end time of each word in the transcript, and the word itself. The audio files are segmented in such a way that the speech recognition can be performed in parallel on different portions of the document. A farm of workstations tied together in a queuing system recognizes each portion of the document. Thus, even if the speech decoder is not real-time, we can still achieve sub real-time throughput. When all the portions are recognized, the results are assembled to create a fully annotated document. Calista yields an error rate of about 20% on a single pass search on the 1998 ARPA HUB4 evaluation corpora (DARPA, 1998) with an average computational load of 6 times real time on a Compaq DS20 EV6 Alpha processor running at 500 MHz. The production system is composed of a farm of 30 Compaq AP500 and 1850 dual Pentium II/III, 450 MHz and 600 MHz, 256 Mb RAM, running Linux 6.0.

When the transcription is available, we replace the speech recognition module with an aligner module whose role is to provide time marks for each word of the input text. We developed a method for aligning transcripts to very long audio segments. This method is robust to occasionally inaccurate transcripts, music insertions, or noisy speech. The precision measured is over 99% of words aligned within a 2 seconds misalignment (Moreno *et al.*, 1998).

### Librarian

The librarian manages the workflow of tasks to be performed to index documents and store meta-data information for each document. Each registered process of the workflow can request to the librarian for work to be performed; this includes such tasks as speech decoding, text/audio alignment or insertion into the index. The librarian is built on an Oracle relational database running on a Compaq DS20 AlphaServer EV6, 2 Gb RAM, running Tru64 Unix 4.0. This centralized model leads to a robust distributed architecture which can scale on demand.

Search for:  Look for:

Show:  Dates:

[New Search](#) · [Search Tips](#)

**Search Result:** 200 matches for your query

Sort by:


Show	Date	Most Relevant Clip
* Speech recognition excerpts do not match audio <u>exactly</u>		
 <b>Fresh Air</b> 58 min	Sep 20, 1999	... washington I'm craig windham a strong <b>earthquake</b> has shaken the capital of taiwan's top billing buildings and knocking out power and phone service come much of the city there been no...
		
<a href="#">Show me more</a>		
 <b>Talk of the Nation</b> 1 hr 34 min	Oct 28, 1999	... of nineteen ninety four might print and I worked viewed as having me that the california and I'm I can read or the exact dates but the <b>earthquake</b> was the exact date...
		
<a href="#">Show me more</a>		
 <b>Sightings on the Radio with Jeff Rense</b> 2 hr 51 min	Oct 11, 1999	... independence of the key to survival school was to have which you're going to lead me to get into the emergency and has and it doesn't really matter whether that emergency is an <b>earthquake</b> ...
		
<a href="#">Show me more</a>		
 <b>Sightings on the Radio with Jeff Rense</b> 2 hr 51 min	Oct 20, 1999	... matter whether that emergency <b>earthquake</b> of what the tornado or a period of employment hungry people are vulnerable to these basic to the independent than survival...
		
<a href="#">Show me more</a>		
 <b>Fresh Air</b> 58 min	Aug 24, 1999	... last week's <b>earthquake</b> but the government continues to reject criticism that it responded to slowly to the crank in the b. b. c.'s chris morris reports from on current as many...
		
<a href="#">Show me more</a>		

Figure 2: The SpeechBot UI

## Indexer


The role of the indexer is not only to index the words that were spoken, but also to be able to retrieve the time when a particular word or sentence occurred. The indexer module catalogs the documents based on the transcription produced by the speech decoder using a modified version of the AltaVista query engine (Burrows, 1998). While the original version of the indexer returns the document containing the query words, our modified version returns multiple hits per documents, one hit for every location in the document that matches the query.

Since documents can be arbitrarily long, our system provides random access within documents. Documents are segmented into fixed length sub-documents, or *clips* (20 seconds long). In order to sort the documents by relevance, we use the term frequency / inverse document frequency (tf/idf) metric

(Salton and McGill, 1983), with each match term multiplied by a position-dependant factor. In addition to the transcription of the audio, we also index the meta-data if any is available. Both the librarian and the indexer are based on the system described in (Eberman *et al.*, 1999).

### User Interface

The Web server returns up to 20 pages of up to 10 documents each sorted by relevance (see Figure 2). For each document, we display a section of the transcription with the query word/s highlighted. This transcription is typically produced by the speech decoder and may contain errors. In a few cases, the text has been provided with the audio so the aligned text is displayed. Links to the original document (*Play Program*) or the audio clip (*Play Clip*) corresponding to the displayed transcription are provided. All the clips of interest can be displayed along a timeline when clicking on *Show me more* (see Figure 3). By default, this page shows 30 seconds of the current selected clip. The window can be enlarged up to 2 minutes. Clips can be selected either directly from the timeline or from a list of clips sorted by relevance (pull-down menu at the bottom).

PLAY PROGRAM

**Fresh Air**  
58 min

Sep 20, 1999

[Find within this program](#) · [New Search](#) · [Search Tips](#)

  
You are here  
Clip 1 of 2

Display 30 seconds

... in from national public radio news in washington I'm craig windham a strong **earthquake** has shaken the capital of taiwan's top billing buildings and knocking out power and phone service come much of the city there been no reports yet of any injuries in taipei but the u. s. geological survey is **earthquake** centered to the quake had a magnitude of seven point six the timber was followed by several aftershocks president clinton has announced more federal assistance for people hard hit by hurricane force to help or range from food stamps to low interest loans mr. clinton toured some of the flooded ...

PLAY CLIP

washington I'm craig windham a strong earthquake has shaken the capital of taiw

Find within this program

Any of these words

FIND

Figure 3: Browsing the clips of a document.

### 3. Performance Analysis

Although we are ultimately interested in the quality of information retrieval, we will examine first the performance of the speech recognition system.

#### 3.1 Speech Recognition Performance

The speech recognition system has been tested on randomly selected segments from several popular radio shows: *Coast to Coast* with Art Bell, Dr Laura Schlessinger, Ed Tyll, Rick Emerson, and the Motley Fool Radio Show. Four 15-minute segments were selected from each of five shows for word error rate analysis. The majority of the audio streams are encoded with the 6.4 kbps RealAudio codec. A few audio streams are encoded at higher rates. After download, the audio is stored in a wav file sampled at 8 kHz. Acoustic conditions vary as shows may have telephone conversations, commercials, several people talking simultaneously, or music in the background. The selected segments are transcribed manually, and the transcription is used to estimate the Word Error Rate (WER). Here  $WER = (\text{substitutions} + \text{insertions} + \text{deletions}) / \text{total number of words}$ .

We break each 15-minute test segment into smaller utterances that are more easily handled by the recognizer. We use a language model trained using the DARPA broadcast news Hub-4 1998 text corpora. It contains a vocabulary of 64,000 words which corresponds to 4 million bigrams and 15 million trigrams in the language model.

We built two sets of acoustic models, one from the original training set and the other from the same corpus, but after an audio compression/decompression transformation. This transformation simulates the quality of audio downloaded from the Internet.

The first set of acoustic models was trained on 100 hours of the Broadcast News corpus provided by LDC at its original 16 kHz sampling rate with recording studio acoustic quality. When using these models the test data must be up-sampled to 16 kHz. The second set of models was trained on the same training corpus which had been encoded using the 6.5 RealAudio Kbps codec, and then decoded to a sampling rate of 8 kHz. This encoding/decoding operation was done to reduce the acoustic mismatch between the training corpora and the testing corpora.

Show	Segment 1	Segment 2	Segment 3	Segment 4
Art Bell	50.5 %	46.3 %	51.2 %	44.5 %
Dr Laura	51.2 %	47.4 %	52.7 %	59.2 %
Ed Tyll	62.3 %	54.0 %	49.9 %	47.0 %
Rick Emerson	48.2 %	51.5 %	53.1 %	56.4 %
Motley Fool	44.1 %	38.8 %	47.2 %	43.5 %

Figure 4: Speech recognition WER, 8kHz RealAudio models,  
16 Gaussian mixture components per HMM state.

On the test set we observed an improvement in the average word error rate from 60.5% for the first set of acoustic models to 49.6% when the 8 kHz Real Audio encoded/decoded models were used. These results are for a system with 16 Gaussian mixture components per HMM state. Figure 4 shows full results for the 8 kHz Real Audio system.

The Calista speech recognizer can be adjusted to trade off processing time for accuracy within certain limits. Ranges between 6 and 30 times longer than real time on a Pentium II, 450 MHz processor under Linux give good accuracy/processing time tradeoffs for a production system. The dominant parameter is the 'beam' of the recognition search. During recognition, hypotheses with likelihood less than the beam are discarded, reducing the number of calculations required. Figure 5 shows the performance of the recognition system as this factor is adjusted. In production mode we use 8 Gaussian mixture components per state and a beam of  $1e-58$ . From Figure 5, the recognizer then runs at around 13 times real time. Because the speech recognizer is parallelized when running on archived

data, and assuming an average of 13 times real-time, we are able to process approximately 110 hours of audio per day on a farm with 60 processors.

Number of Gaussians	Beam	Error Rate	Average times real time
16	1e-64	49.6%	23.2
16	1e-58	51.5%	18.5
8	1e-64	51.9%	18.0
8	1e-60	53.0%	14.5
<b>8</b>	<b>1e-58</b>	<b>53.9%</b>	<b>13.0</b>

Figure 5: Speech recognition WER and speed as a function of beam and number of Gaussian mixture components per HMM state (8kHz Real Audio Models)

### 3.2. Information Retrieval Performance

The retrieval performance was evaluated on the public site with static content. We had independent and unbiased testers evaluating a set of queries selected from the list of the most frequently submitted queries.

Since the shows that we indexed were often hours long and did not have topic boundaries, we had to devise some way of subdividing shows. In response to the user's query, we first rank all the returned shows and then divide each show into chunks of equal size (or *clips*), and rank them within each show. Each of the retrieved clips has a link back to the corresponding part of the audio show on the original broadcast site. The advantage of ranking clips within the shows as well as ranking complete shows is that the user ideally sees only the relevant parts of the longer documents. This also makes the process of searching for relevant information much easier. Due to the sequential nature of audio data, it is impossible to skim longer documents in search of relevant information, as with textual search. Our solution allows the user to go directly to the relevant parts of the retrieved audio shows.

We evaluated retrieval performance on an index made from a collection of 22 programs, containing about 3,111 shows corresponding to roughly 3,000 hours of audio downloaded off the Web. We used 40 queries, measuring the precision of the top 20 retrieved whole shows. The queries were selected from the top 100 queries submitted to the public site since December 1<sup>st</sup>, 1999 (SpeechBot, 1999). The words were selected such that they cover a large variety of topics, varying length of words (phoneme-wise) and varying types of words such as acronyms or proper nouns. The queries were either nouns or noun phrases with a maximum length of 3 words. Example queries are "bill clinton," "internet," and "y2k". All of the words used for evaluation were in the vocabulary of the speech recognition engine. The top 20 documents for judging relevance were selected based on the assumption that typical users only tend to look at the first couple of pages of the retrieved results (Silverstein *et al.*, 1998).

In order to judge relevance of the retrieved documents, we used a "concept hit" approach, i.e. the document is considered relevant if and only if the concept described in the query is present in the retrieved document. For example, for a document to be relevant to the query "aliens," it had to refer to aliens from the outer space, not immigrants. Relevant concepts were provided to the judges along with queries. For a document to be relevant to the query "y2k," it was enough to mention "the millennium bug." We had three independent human judges who were given the same instructions as to the notion of relevance for each query and the same set of queries. Each judge was given the ability to look at the transcripts of the retrieved documents as well as to listen to the respective parts of the audio show. The judges were not allowed to discuss relevance of any document among them.

To evaluate the retrieval results, we used a standard average precision metric (Salton, 1989):

$$P = \frac{N_r}{T_r}$$

where  $P$  is precision,  $N_r$  is the number of relevant documents retrieved, and  $T_r$  is the total number of documents retrieved. For these 40 queries, the average precision of the top 20 documents is 69%. A previous experiment done on a smaller index, and with words chosen from the transcription of the shows gave similar results. The average precision for the retrieved 200-word chunks was 72.67% while for the top 5 retrieved whole shows it was 65%. Note that, with the current size of the index, none of the queries returned less than 20 documents.

The retrieval performance of the system is better than expected considering the accuracy of the speech decoder. There are several reasons why. First the query words are often repeated several times during a show and are thus more likely to be recognized. Second, the keywords tend to be longer than stop words so the speech recognition search is more constrained and tends to perform better on these words.

Retrieval errors are due to one of two main reasons. First, insertion or substitution recognition errors cause query words to appear erroneously in the transcripts. This kind of error represents roughly half of the cases of the appearance of non-relevant documents

The other main reason for a non-relevant document to be retrieved is the so-called "out-of-context mentioning" of the query words. For example, a commercial in the middle of a program can briefly mention the query words, but not mention the query subject otherwise. For example, a country name, a subject of one of our queries, was mentioned in one airline commercial. Note however that some commercials were considered relevant if they truly dealt with the subject matter (e.g., a commercial for a wireless service provider for PDA's was considered relevant to the query "wireless" by one of our judges).

"Out-of-context" referrals can also arise because of the inherent ambiguity of the query words. For example, the query "AIDS" returned many documents which talked about "aids" meaning "helps" rather than a disease. One way to approach this problem would be to introduce a notion of subject popularity; a good example is the popularity-based search engine Google (Google, 2000). Yet another example of "out-of-context" referrals is a "brief mentioning of the subject." For instance, for a query "IBM" a talk show was returned where one of the guests was a former IBM employee. This fact was mentioned twice without any further discussion related to the company. With the help of IR techniques such as query expansion and others, we will introduce bias into our ranking scheme in order to overcome these problems (Singhal, 1999; Jourlin *et al.*, 1999).

In some cases, when we used multiple word queries, the query words were distributed among retrieved clips within the same document. If the query was a phrase, an "exact match" option would be appropriate here, but we assumed that typical users don't use the advanced options, but instead rely on the defaults. We will deal with this problem by enforcing a proximity bias in our ranking system. In fact, this bias is already a part of our ranking scheme, but further work is needed to determine the various importance weights.

In an attempt to improve results, we tried query stemming using a Porter stemmer (Porter, 1980; Frakes & Cox, 1986). That is, we assumed that while some words might have not been recognized correctly, their stems were. The improvement was very small and not worth the overhead of stemming each query. We additionally noted that in some cases (not in our reported test results), unsuccessful queries were due to obvious recognition failures such as when the query word is not in the dictionary.

At the time of writing, no evaluation has been performed with the TREC-SDR test set. This evaluation would be of interest and is slated for future work. However, note that the test we ran on real data (both audio content, and query-wise) is six times bigger than the TREC-8 audio set (TREC, 1999). The diversity of audio condition and content is also higher. Finally, the quality of the audio (decompressed after download) makes this experiment unique.



#### 4. Conclusion and future work

SpeechBot is the first audio indexing and retrieval system for the Web. It incorporates speech recognition technology and so is able to operate when no transcription is available. The current version of the system is capable of indexing up to 800 hours of audio a week, and is configured in such a way that it can easily scale up on demand. It has given us a unique opportunity to test a spoken data retrieval system on a large scale. Experiments show acceptable retrieval accuracy despite high recognition error rates. This suggests that indexing of audio documents on the Web is feasible given current level of recognition and retrieval technologies.

There is still room for many improvements. Speech recognition accuracy can be increased by providing specialized audio models, vocabulary, pronunciation dictionaries, and language models. To improve IR accuracy, we are investigating different relevance ranking algorithms, and alternate ways of indexing the output of the speech recognizer, as well as query and document expansion. We also plan to improve the user interface since random access to multimedia documents raises interesting new questions about the efficiency of the current design. Finally, we plan to port the speech recognition engine to other languages, most likely French and Spanish.

#### Acknowledgments

The SpeechBot project has benefited from the insights and work of many people. We wish to especially acknowledge Ron Gentile, and Angel Chang. System design and engineering work have been done by R. Paul Johnson, Chris Weikart at CRL, and Katrina Maffey, Blair Fidler, Matthew Moores, Mostafa AbuShaaban at SEA Australia. We acknowledge Catherine Warner of WRL for the help with the UI design. Thanks to Andrew Leonard, we have a big and rich index of audio content. Adam Bryant, and Kamau Wanguhu from CRL provided world-class system support. Finally we would like to thank the ASE testing team: Frank Bomba, Gene Preble, and John Axon for their support for testing and evaluating the system.

#### References

- Abberley D., Cook G., Renals S. & Robinson T. (1999). Retrieval of Broadcast News Documents with the THISL System. In *Proceedings of the Seventh Text Retrieval Conference (TREC-8)*. Gaithersburg MD, USA.
- Burrows M. (1998). Method for Indexing Information of a Database. U.S. Patent 5,745,899.
- DARPA. (1998). Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop. Landowne, VA.
- Eberman B., Fidler B., Iannucci R.A., Joerg C., Kontothanassis L., Kovalcin D.E., Moreno P., Swain M.J., & Van Thong J-M. (1999). Indexing Multimedia for the Internet. In *Visual Information and Information Systems*. D. P. Huijsmans and Arnold W.M. Smeulders (Eds.) Springer-Verlag.
- Frakes B. & Cox C. (1986) The Porter stemming algorithm. Software available from <http://www.cs.jhu.edu/~weiss/stem.c>.
- Garfalo J., Vorhees E., Auzanne C., Stanford V. & Lund B. (1998). Spoken Document Retrieval Track Overview and Results. In *The Seventh Text Retrieval Conference (TREC-7), NIST Special Publication* (500-242).
- Google (2000). <http://www.google.com>
- Johnson S. E., Jourlin P., Moore G. L., Sparck Jones K. & Woodland P.C. (1999). The Cambridge University Spoken Document Retrieval System. In *Proceedings of the IEEE International Conference On Acoustics, Speech, and Signal Processing*.
- Jourlin P., Johnson S.E., Sparck Jones K, Woodland P.C. (1999). General Query Expansion Techniques for Spoken Document Retrieval. In *Proceedings of the ESCA Workshop on Extraction Information from Spoken Audio*.

- Kontothanassis L., Joerg C., Swain M. J., Eberman B., & Iannucci R.A. (1999). Design, Implementation, and Analysis of a Multimedia Indexing and Delivery Server. Technical report CRL 99/2, Compaq Cambridge Research Laboratory.
- Moreno P.J., Joerg C., Van Thong J-M. & Glickman O. (1998). A Recursive Algorithm for the Forced Alignment of Very Long Audio Segments. In *Proceedings of ICSLP'98*.
- Porter M. (1980). An Algorithm for Suffix Stripping. *Program* 14 (3), (pp. 130-137).
- Salton G., & McGill M. J. (1983). Introduction to Modern Information Retrieval. McGraw-Hill.
- Salton G. (1989). Automatic Text Processing: The transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, Reading, Massachusetts.
- Silverstein C., Henzinger M., Marais H., Moricz M. (1998). Analysis of a Very Large AltaVista Query Log. SRC Technical Note 1998-014. October 1998.
- Singhal A., Pereira F. (1999). Document Expansion for Speech Retrieval. *ACM SIGIR'99*.
- SpeechBot (1999). <http://www.compaq.com/speechbot>
- TREC (1999). [http://www.itl.nist.gov/iaui/894.01/sdr99/doc/sdr99\\_spec.htm](http://www.itl.nist.gov/iaui/894.01/sdr99/doc/sdr99_spec.htm)
- Wactlar H.D., Hauptmann A.G. & Witbrock M.J. (1996). Informedia: News-on-Demand Experiments in Speech Recognition. In *Proceedings of ARPA Speech Recognition Workshop*, Harriman NY, USA.
- Broadcast (1999). <http://www.broadcast.com/radio/>
- Witbrock M.J. & Hauptmann A.G. (1997). Speech Recognition and Information Retrieval: Experiments in Retrieving Spoken Documents. In *Proceedings of the DARPA Speech Recognition Workshop*.
- Witbrock M. (1999). SearchEngine Meeting, Search Engines and Beyond. Boston. <http://www.infonortics.com/searchengines/boston1999/witbrock/index.htm>, Lycos, Waltham, MA, USA.