COMPAQ

Technical Whitepaper

The First Internet Site for Content-Based Indexing of Streaming Spoken Audio

www.compaq.com/speechbot

Compaq Computer Corporation
Cambridge Research Laboratory
One Cambridge Center
Cambridge, MA, 02142

1. Overview

SpeechBot (http://www.compaq.com/speechbot) is the first Internet search site for indexing streaming spoken audio on the web. Unlike previous attempts to index spoken audio on the Web, which have relied on either adjacent text, metadata, or hand supplied transcripts and close captions, SpeechBot uses automatic speech recognition technology to transcribe and index documents that do not have transcripts or other content information. The use of speech recognition permits the efficient and cost-effective indexing of thousands of hours of audio content, which were previously inaccessible. Because of this indexing, SpeechBot allows users to quickly search for relevant content in long audio documents and yields a high precision on first page-retrieved items.

SpeechBot indexes streaming media files based on their content, much as conventional search sites index ordinary Web pages by their text content. Like conventional search sites, SpeechBot does not store or serve the multimedia files themselves, but rather provides users with links. SpeechBot's current index has over 3200 shows, 3500 hours of audio and 20 million words. The index is continually updated using SpeechBot's highly scalable architecture.

The SpeechBot site was designed as a demonstration vehicle for Compaq Corporate Research's work on multimedia indexing for the Internet. This technology includes workflow systems for crawling, downloading, transcoding, and indexing multimedia files, several technologies (including speech recognition) for analyzing and annotating multimedia content, and search site technology to handle user queries and serve appropriate Web pages.

2. Introduction and Motivation

The amount of audio and video content available on the Internet has increased dramatically over the past few years. One of the largest segments of this content is the growing number of spoken audio documents archived on the Web. One part of this material consists of studio-produced broadcasts, such *as PBS's News Hour* for which there are transcripts and summaries, which can be used as a basis for indexing. However, another rapidly growing segment consists of talk shows, interviews, lectures, speeches and talks. Table 1 lists a number of popular talk-radio shows archived on the Web. These few shows alone account for over 100 hours of new audio added to the Web per week.

Show	Hours/week	Content Provider (rank)
Art Bell	35	Artbell.com
Rick Emerson	20	Broadcast.com
Jeff Rense	18	Broadcast.com
Dr. Laura	15	Drlaura.com
The Connection	10	Wbur.org
Diane Rhem	10	Wamu.org
Fresh Air	5	Whyy.org
Motley Fool	3	Broadcast.com
Car Talk	1	Cars.com
InternetNews radio	1	internetnews.com

Table 1: Rate of new content production for talk radio content archived on the internet.

For most of this material there are no transcripts available and therefore no simple means for indexing their content. For example, except for some carefully curated archives such as National Public Radio site (NPR), most archives of radio shows on the Web are simple lists of links to long audio files (see [Web1]). 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.

The goal of SpeechBot is to address this problem using automatic speech recognition to produce transcripts for audio when they are not otherwise available. Although considerable progress has been made in large-vocabulary speech recognition, due in part to research sponsored by DARPA, this technology is far from perfect, particularly when applied to the audio that has been degraded by the coding and compression schemes which enable the delivery of streaming audio over the Internet. However, recent research in spoken document retrieval has been shown to be remarkably resilient to word error rate. A recent study [Wit99] shows that a word error rate of 30% reduces recall by 4%, and a word error rate of 50% reduces it by only 10% (see also [Gar98]). There are several explanations for these numbers. Query keywords tend to occur multiple times in relevant documents; if a word is missed once by the recognizer, it may still be successfully recognized other times it appears in the document. Also, many queries consist of multiple keywords, increasing the odds that as least a subset will be successfully recognized. Finally, it is likely that query keywords are longer and more acousticly distinct than "stop words", ('the','a','in', etc.) and may have a lower recognition error rate.

The SpeechBot site currently indexes over 3500 hours of audio representing over 30 million spoken words. This is only a snapshot; additional content is continuously added as it appears online. Supporting this continuous update is a scalable architecture developed to crawl, download, transcode, transcribe, and index Web multimedia in a variety of formats. This system is an extension of previous work at Compaq's Cambridge Research Laboratory in multimedia search and indexing, which was successfully transferred to AltaVista and is used in its Media Finder.

The remainder of this document describes various aspects of SpeechBot in greater detail. Section 3 gives an overview of speech recognition for streaming spoken audio including some experimental results on retrieval and recognition accuracy. Section 4 describes the site itself, the user interface and some of the key features. Section 5 briefly describes the workflow and architecture supporting SpeechBot. Finally, related research at Compaq's Cambridge Research Lab is outlined in Section 6.

3. Speech Recognition for Indexing Spoken Document

Large-vocabulary automatic speech recognition (ASR) has been the object of extensive research and development over the last decade, improving recognition accuracy and computational efficiency. The accuracy of modern ASR systems, in practice, depends of several factors including: the quality of the audio, the variety of speakers, the style of speaking, and the content of the speech. In many ways, the spoken audio content found on the Web is difficult across all of these conditions. The audio has been compressed and processed for efficient storage and streaming which degrades the quality of the speech. Furthermore, Web content contains a large variety of speakers with various accents, speaking rates, and conversational styles. Finally, this content covers a wide array of topics make creation of language models, an important component of ASR systems, particularly challenging. As a result of these factors, recognition accuracy for Web audio transcription is far below that possible for more constrained tasks, such as single-speaker dictation. However, recent research in information retrieval has suggested that current recognition accuracies are sufficient for effective indexing of spoken documents [Wit97].

3.1. Speech Recognition Engine

SpeechBot currently uses a large-vocabulary speech recognition engine developed at Compaq's Cambridge Research Lab. This technology behind this engine is typical of current research systems; Hidden Markov Models for acoustic modeling and backoff trigrams for language modeling [Rab93]. Figure 1 shows the internal processing stages of the recognizer.

The audio signal is first processed into a sequence of 'feature vectors', numbers which capture important properties about the signal. One feature vector is produced for each 10ms 'frame' of input speech. These feature vectors are then scored against a set of acoustic models which represent the various sounds that make up the speech of a given language. These score vectors, one of each frame, are then grouped into higher and higher levels of organization. First into phonemes, then, based on the specified lexicon and language model, into a sequence of words.

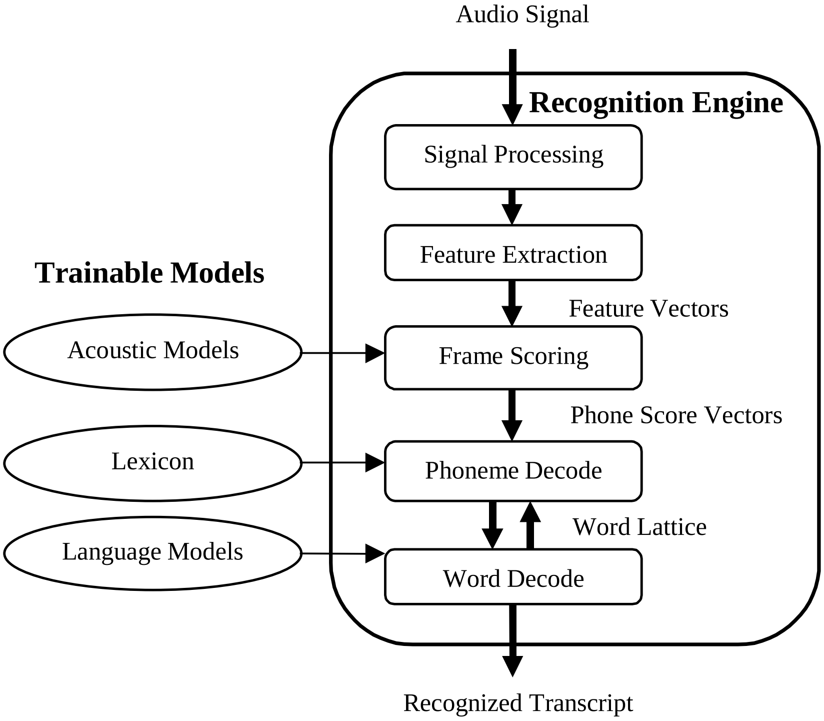


Figure 1: **Speech Recognition Internal Architecture**

One of the key properties of this speech recognition technology is that it is trainable. Its behavior is controlled by three components; the acoustic models representing each sound, the lexicon specifying the words to be recognized and their pronunciation, and the language model capturing (at some level) the linguistic properties of language. To achieve optimum recognition performance on any task, each of these components needs to be trained for the particular characteristics of that task.

3.2. Recognition Accuracy on Web Audio

The initial accuracy of the speech recognition has been measured on RealAudio compressed Web audio data. For these experiments, segments were selected from four popular talk radio shows archived on the Web: 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 wav format sampled at 8 kHz. Acoustic conditions vary both across and within segments as shows can have telephone conversations, commercials, several people talking simultaneously, or music in the background.

Two alternative sets of acoustic models were evaluated. The first set was trained on 100 hours of the Broadcast News corpus provided by the Linguistic Data Consortium (www.ldc.upenn.edu) at the original 16 kHz sampling rate. When using these models the test data must be up-sampled to 16 kHz. The second set of models was trained on a version of the LDC training corpus, which had been encoded using the RealAudio 6.5 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. Research has shown that better accuracies are achieved when the training conditions are closely matched to those of actual operation. All experiments used a trigram language model trained on text corpora from the LDC. This language model supports a vocabulary of 64,000 words, and consists of 4 million bigrams and 15 million trigrams.

As expected, we observed an improvement in the average word error rate from 60.5% for the 16kHz models to 49.6% for the 8 kHz Real Audio encoded/decoded models. These results are for a system with 16 Gaussian mixture components per HMM state. Table 2 shows full results for the system, 8 kHz Real Audio trained models.

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 %

Table 2: Speech recognition accuracy, 8kHz RealAudio models, 16 Gaussian mixture components per HMM state.

The 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 In production, the system runs at an average of 13x real time. Segments with music, noise, or long silences tend to take longer while segments containing clean speech run faster. Because the speech recognition process is parallelized in the SpeechBot architecture, we can process around 110 hours of audio a day on a farm with 60 machines.

3.3. Information Retrieval Experiments

Before proceeding with the development of SpeechBot, we investigated the retrieval precision that could be expected from indexing documents based on automatically recognized transcripts. For this purpose, we used a collection of episodes of the Art Bell show containing about 100 hours of audio. After obtaining the automatic transcripts from the recognizer, indexed them and provided a querying engine on top of it.

Since the shows that we indexed were often hours long, did not have topic boundaries, and we had no ranking algorithms suitable for audio content, we had to devise some way of subdividing shows into subdocument or 'clips'. Our approach was to divide each show into 200-word chunks based on the output of the speech recognizer, and to then rank these chunks in response to the user's query. Each of the retrieved subdocuments has a link back to the corresponding part of the audio show on the original broadcast site. We also performed a set of experiments where entire shows were returned in response to the user's query. The advantage of ranking only small subparts of the shows is that the user ideally sees only the relevant parts of the longer documents. This also makes the process of performing relevance judgments much easier.

The index was evaluated on a set of short queries to obtain the precision of the top 20 retrieved clips. The queries were selected based on the show summaries from the show's Web site. Example queries are "tree sitter," "global warming," and "Kennedy assassination". The queries were either nouns or noun phrases with a maximum length of 3 words. The top 20 clips 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 [Sil98]. This analysis will be refined with queries reflecting the actual usage of the system when we have collected more data.

In order to judge relevance of the retrieved documents, we used a "concept hit" approach, i.e. the document was considered relevant if and only if the concept described in the query was 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. For a document to be relevant to the query "y2k," it was enough to mention "the millennium bug." We had two 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 between themselves. Despite this, they had identical judgements about what documents are considered to be relevant

To evaluate the retrieval results, we used a slightly modified standard average precision metric [Sal89].

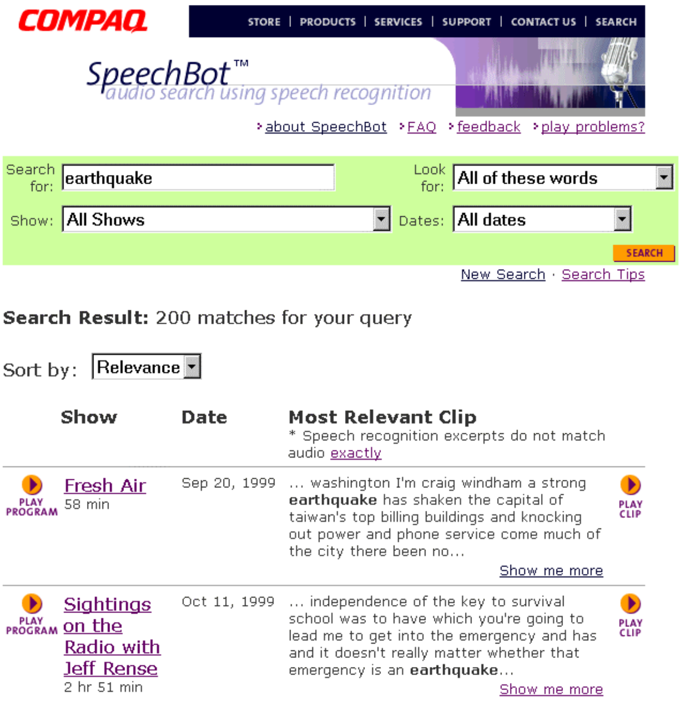
$$\text{Precision} = \frac{\text{Relevant DocumentsRetrieved}}{\min(20, \text{Total Number Of Documents Retrieved})}$$

For these 25 queries, the average precision of the top 20 retrieved 200-word chunks was 72.67. It is important to note that a number of queries retrieved less than 20 documents. Therefore, precision figures for those queries were higher.

4. SpeechBot Search Site

The SpeechBot site is modeled after conventional Internet search sites such as AltaVista. Like these sites, it does not maintain and server any media content directly. Rather, it indexes multiple media serving Web sites and returns to the user links to these sites in response to search queries.

The entry page of the SpeechBot site looks familiar to search engine users. There are boxes to enter query keywords and specify query types. In addition, there are some options particular to this class of content. The user can restrict searches by air date (for archived radio shows) or show name. The entry page gives a list of the programs currently indexed and links back to the content host sites.



COMPAQ
STORE | PRODUCTS | SERVICES | SUPPORT | CONTACT US | SEARCH
SpeechBot™
audio search using speech recognition
about SpeechBot
FAQ
feedback
play problems?
SEARCH
New Search · Search Tips

Search for: earthquake

Look for: All of these words

Show: All Shows

Dates: All dates

Search Result: 200 matches for your query

Sort by: Relevance

	Show	Date	Most Relevant Clip * Speech recognition excerpts do not match audio exactly	
PLAY PROGRAM	Fresh Air 58 min	Sep 20, 1999	... 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... Show me more	PLAY CLIP
PLAY PROGRAM	Sightings on the Radio with Jeff Rense 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 **earthquake**... Show me more	PLAY CLIP

Figure 2: SpeechBot Response Page

Figure 2 shows a screenshot of the SpeechBot site user interface. In this case the user has entered a query and gotten back a page of responses. For each response, the interface presents the show title with a link to the content provider, and the date, and length of the full audio program. In addition, it displays a brief section of the transcript around the matched query words, with the match highlighted in bold. While the transcripts produced by the automatic speech recognition have a considerable number of errors, they are often sufficient for determining if the returned clip is worth further investigation. To play the original audio, the user has two options. There is a link to the matched audio document (often several hours long) for the user to listen to in its entirety, and a link to the specific clip matching the query. This enables the user to start listening in the vicinity of the match.

If desired, the user can request more information about the match. This link brings up a page containing a time line representing the entire audio document with marking indicating the location of query hits. The user can navigate around the document, examining each hit. Additionally, there are provisions for showing more of the transcription in the vicinity of the query hit.

5. Scalable Architecture for Multimedia Indexing

The basic information flow through the SpeechBot architecture is shown in Figure 3. A multimedia document to be indexed is first downloaded from its host location and transcoded into a common format for downstream processing., for audio documents, this is 16-bit PCM (.wav) format. At the same time metadata from the download is collected and stored in a database. This metadata includes show title, air date, duration, and copyright information, as well as the links back to the original host. The transcoded audio is further processed, if necessary, to convert to the 8kHz sampling rate assumed by the speech recognition. Automatic speech recognition is then used to generate a transcript which is the used to index the document. The transcript produced by the speech recognition is time-stamped at the word level so the index can point to the precise location of each word in each document. Once transcribed and indexed, the original audio and all processed derivatives are deleted from the system.

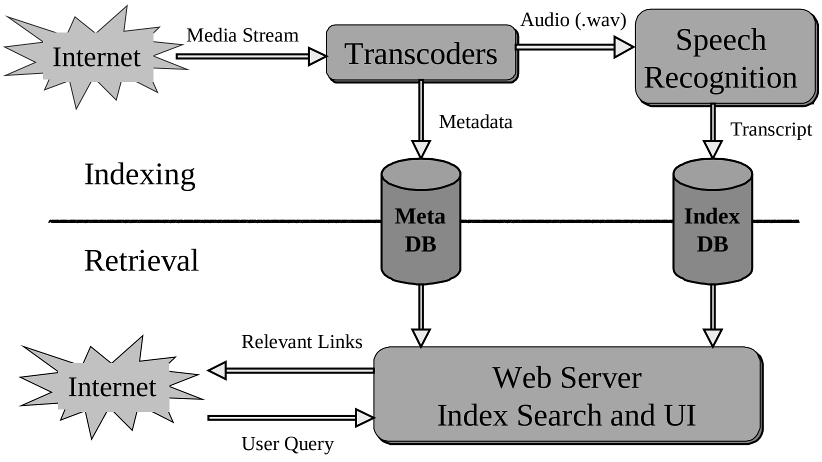


Figure 3: SpeechBot Data Flow

On the retrieval side, user queries are looked up in the index for relevant document ids. Any hits are looked up in the metadata database to retrieve the associated information. Finally, the collected metadata, links and transcript extracts for relevant hits are composed into an HTML page and returned to the user.

Supporting this dataflow is a scalable architecture designed to handle the increasing amount of multimedia content. The download and transcoding processes and handled by a collection of ‘worker bees’, workstations under the control of a ‘queen bee,’ which manages job dispatching. Since Web audio in streaming data formats (the majority of audio currently indexed by SpeechBot) must be downloaded in real time, several jobs are run simultaneously on each worker bee. The number of jobs that can be run is limited primarily by the storage on these machines, as the transcoded audio files are quite large. The worker bees are instantiated as Compaq Professional Workstation running the Window NT operating system.

The speech recognition is performed on a farm of Compaq ProLiant Servers running Linux. In order to improve throughput and robustness, each long audio file is broken into small segments for recognition. This insures that if a particular recognition process fails (due to perhaps system failure or network congestion), only a small amount of work is lost and the segment can be easily reprocessed. Each audio segment spawns a recognition task on one of the Linux farm machines. When all recognition tasks associated with a given document are complete, the final transcript is assembled from the partial transcripts produced from each of the segments. This transcript is then inserted into the index database.

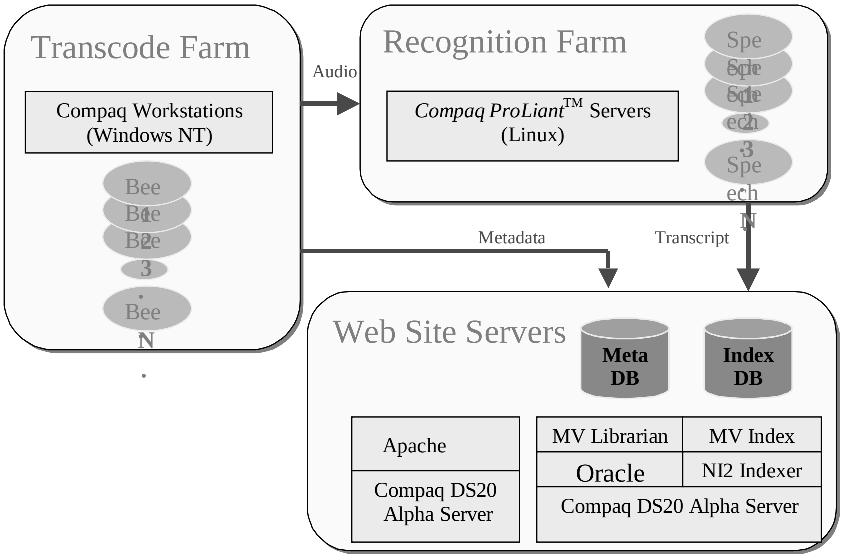


Figure 4: SpeechBot System Architecture

The two databases and the query server are supported by Compaq DS20 Alpha Servers running Tru64 UNIX. The metadata is maintained by a Librarian system built on top of Oracle. The index database uses the NI2 [Bur98] indexing engine, which was first developed by Compaq Corporate Research for AltaVista. Web serving is done by an Apache server running on one of the DS20's. This architecture was designed from the start to be heterogeneous to allow flexibility in processor and operating system choice. Communication between the various system elements is mainly through remote procedure calls implemented as cgi-bin requests.

6. Related Research at CRL

Compaq's Cambridge Research Laboratory (CRL) has been conducting research into indexing and searching multimedia on the Internet for several years. This work has produced technologies in three areas: multimedia crawling, multimedia annotation, and indexing long multimedia documents. The first and third of these have been incorporated into SpeechBot and are described above. Multimedia annotation refers to a collection of technologies, which add information to raw media files. This information can then be used to better index, retrieve and present multimedia documents. Some examples of annotation technologies under development at CRL are:

Automatic Keyframe Selection: Given a video document of significant length, find a keyframe or small set of keyframes which best represent the video. These keyframes are used in presented query hits to users as a further guide to relevance.

Speech Alignment: [Mor98] In some instances, some form of transcript is available for long multimedia documents. However, these transcripts are not time-stamped making selection of relevant portions of long documents impossible. The alignment technology developed at CRL efficient matches the audio track with the given transcript and produces timestamps for each word in the transcript.

Face and Skin Detection: In indexing and classifying the large number of photographs on the Web, the detection of skin and faces is important for several applications, including “family-friendly” filtering.

Audio Classification: [Mor2000] This work attempts to automatically classify unknown audio files with varying degrees of specificity from simple speech/music/other distinctions to more fine-grained classifications (for example, music by primary instrument).

Music Keyframing: [Logan2000] Select a representative 10 second portion from a longer song. This can be used to let users sample content before committing to download over slow links.

7. Summary

SpeechBot is an Internet search site (www.compaq.com/speechbot) for finding spoken media on the Web. This site indexes streaming media files based on their content, much as conventional search site index ordinary Web pages by their text content. Like conventional search sites, SpeechBot does not store or serve the multimedia files themselves, but rather provides users with links.

Previous attempts to index spoken audio on the Web have relied on either adjacent text, metadata, or hand supplied transcripts and close captions. SpeechBot uses automatic speech recognition technology to transcribe and index documents for which this information is unavailable. The use of speech recognition permits the efficient and cost-effective indexing of thousands of hours of audio content, which were previously inaccessible, except by random listening.

The SpeechBot site was designed as a demonstration vehicle for CRL’s research in multimedia indexing for the Internet. This technology includes workflow systems for crawling, downloading, Transcoding, and indexing multimedia files, several technologies (including speech recognition) for analyzing and annotating multimedia content, and search site technology to handle user queries and serve appropriate Web pages.

Compaq’s Cambridge Research Laboratory plans to continue research to improve the basic technology behind SpeechBot, such as automatic speech recognition and relevance ranking for audio documents. New content and features may also be added as new technologies become available. We hope that SpeechBot will be useful and valuable both to Internet users, seeking interesting content, and content hosts seeking a wider audience, as well as providing Compaq Research with a source of inspiration for new research directions in this exciting and growing area.

For further editorial information, contact:

Compaq Computer Corporation Eileen Quinn 408-285-9272 eileen.quinn@compaq.com

8. References and Bibliography

[Abb99] D. Abberley, G. Cook, S. Renals and T. Robinson, *Retrieval of Broadcast News Documents with the THISL System,* In the proceedings of the Seventh Text Retrieval Conference (TREC-8), Gaithersburg MD USA, to appear (1999).

[Bur98] Mike Burrows, *Method for Indexing Information of a Database.* U.S. Patent 5,745,899. April 1998.

[Ebe99] B. Eberman, B. Fidler, R. A. Iannucci, C. Joerg, L. Kontothanassis, D. E. Kovalcin, P. Moreno, M. J. Swain, and J-M. Van Thong. *Indexing Multimedia for the Internet. In Visual Information and Information Systems*, D. P. Huijsmans and Arnold W.M. Smeulders (Eds.) Springer-Verlag, 1999.

[Gar98] J. Garfolo, E. Vorhees, C. Auzanne, V. Stanford, and B. Lund. 1998 TREC7 *Spoken Document Retrieval Track Overview and Results*. In NIST Special Publication 500-242: The Seventh Text Retrieval Conference (TREC 7), 1998.

[Joh99] S. E. Johnson, P. Jourlin, G. L. Moore, K. Sparck Jones and P.C. Woodland*. The Cambridge University Spoken Document Retrieval System*. In Proceedings of the IEEE International Conference On Acoustics, Speech, and Signal Processing, 1999.

[Kon99] L. Kontothanassis, C. Joerg, M. J. Swain, B. Eberman, and R. A. Iannucci. *Design, Implementation, and Analysis of a Multimedia Indexing and Delivery Server*. Technical report CRL 99/2, Compaq Cambridge Research Laboratory, 1999.

[Log2000] Beth Logan and Stephen Chu, *Music Summarization Using Key Phrases*, In Proceedings of ICASSP'2000, 2000

[Mor98] Pedro J. Moreno, Chris Joerg, Jean-Manuel Van Thong and Oren Glickman. *A Recursive Algorithm for the Forced Alignment of Very Long Audio Segments.* In Proceedings of ICSLP'98, 1998.

[Mor2000] Pedro J. Moreno and Ryan Rifkin, *Using the Fisher kernel method for audio classification,* In Proceedings of ICASSP'2000, 2000.

[Pro99] Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop. Landowne, VA, 1998.

[Rab93] L. Rabiner and B.H. Juang, Fundamentals of Speech Recognition, Prentice Hall,1993.

[Sal89] G. Salton. Automatic Text Processing: The transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, Reading, Massachusetts, 1989.

[Sil98] C. Silverstein, M. Henzinger, H. Marais, M. Moricz. Analysis of a Very Large AltaVista Query Log. SRC Technical Note 1998-014. October 1998.

[Wac96] H. D. Wactlar, A. G. Hauptmann, and Michael J. Witbrock, Informedia: News-on-Demand Experiments in Speech Recognition. In Proceedings of ARPA Speech Recognition Workshop, Harriman NY, 1996.

[Web1] http://www.broadcast.com/radio/

[Wit97] Michael J. Witbrock and Alexander G. Hauptmann. *Speech Recognition and Information Retrieval: Experiments in Retrieving Spoken Documents*. In Proceedings of the DARPA Speech Recognition Workshop, 1997.

[Wit99] Michael Witbrock, http://www.infonortics.com/searchengines/boston1999/witbrock/index.htm, Lycos 1999.