SPEECHBOT: A CONTENT-BASED SEARCH INDEX FOR MULTIMEDIA ON THE WEB

Pedro J. Moreno, Jean Manuel Van Thong, Beth Logan

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

Blair Fidler, Katrina Maffey, Matthew Moores

Compaq Computer Corporation
Software Engineering Australia
Bond University Research Park
Queensland 4229, Australia

ABSTRACT

As the Web transforms from a text only environment into a more multimedia rich medium the need arises to perform searches based on the multimedia content. In this paper we present an audio and video search engine to tackle this problem. The engine uses speech recognition technology to index spoken audio and video files from the World Wide Web when no transcriptions are available. If transcriptions (even imperfect ones) are available we can also take advantage of them to improve the recognition process.

SpeechBot indexes several thousand 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. 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 (85%).

1. INTRODUCTION

As the magnitude and use of multimedia content on the web grows, in particular large collections of streamed audio and video files, efficient ways to automatically find the relevant segments in these multimedia streams are necessary. Unfortunately, traditional Web search engines are often limited to text and image indexing and many multimedia documents, video and audio, are thus excluded from classical retrieval systems. Even those systems that do allow searches of multimedia content, like *AltaVista* multimedia search and *Lycos MP3* search, only allow searches based on data such as the multimedia file name, nearby text on the web page containing the file, and meta-data embedded in the file such as title and author. Clearly these systems do no perform any detailed analysis of the multimedia content.

Most multimedia archives on the Web are simple lists of links to long audio files, sometimes several hours in length[1]. Very often, there is no transcription available and therefore no simple means for indexing their content. Even when a transcription is available often it is not annotated or linked to the relevant points of the multimedia stream. 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.

SpeechBot is not the first system to offer these capabilities. In fact, there have been several studies which had similar goals [1, 2, 3]. 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. Third, our system is designed to scale up on demand. Finally, our search index is available and running on the web[2].

The content currently indexed is popular talk radio, technical and financial news shows and some conference video recordings. These shows are almost entirely speech, and very few of them have associated transcriptions, unlike TV shows that are often closed captioned in the U.S.

The outline of the paper is as follows. In Section 2 we give an overview of the architecture of the system. In Section 3 we present a performance analysis. In Section 4 we describe our usability studies. Finally, in Section 5 we present out conclusions and suggestions for future work.

2. SYSTEM OVERVIEW

SpeechBot is a public search site similar to AltaVista, which indexes audio from public Web sites such as *Broadcast.com*, *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. Figure 1 presents the system architecture.

[1] see for example *http://www.broadcast.com*

[2] *http://www.compaq.com/speechbot*

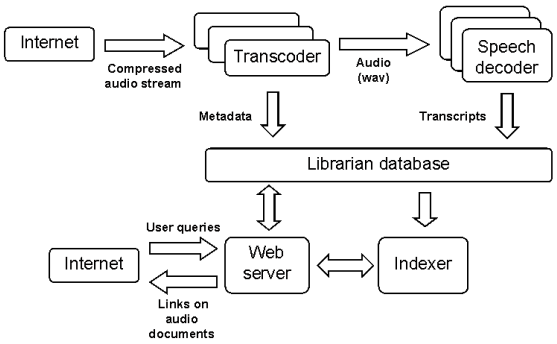

Figure 1: Overall architecture of the system.

2.1. Transcoder

The transcoders fetch and decode video and audio files from the Internet. For each item, they extract the meta-data, download the media documents to a temporary local repository and convert the audio into an uncompressed file format. 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 it is being processed by the system, and to display the result of the query to the user.

2.2. Speech Recognition

In many studies, document retrieval has been shown to be remarkably resilient to the speech recognizer word error rate. A recent study [4] shows that a word error rate of 30% reduces recall by 4%, and a word error rate of 50% reduces it by only 10%. 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.

SpeechBot uses the Calista speech recognizer system. Calista is a large vocabulary continuous speech recognition package using state-of-the-art mixture Gaussian, triphone based, 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 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 [5] with an average computational load of 6 times real time on Compaq workstations running Linux 6.0.

When the transcription is available we replace the speech recognition module with an aligner module. Its role is to provide time marks for each word of the input text. This 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 [6].

2.3. Librarian

The librarian has two main roles. It manages the workflow of tasks carried out by the individual modules, and it stores meta-data and other information required by the user interface.

The component modules often run on remote machines and do not communicate directly. Each process registers itself with the librarian and once registered can make a request for work to perform. This includes such tasks as speech decoding, text to audio alignment or insertion of text into the index. The output of one task is usually the input for another, and the librarian tracks the location of these inputs and outputs.

In addition to storing the meta-data collected by the Transcoder module, the librarian stores 10 second 'clips' of formatted text for each document. It maintains a mapping between word locations from the index, corresponding text clips, and the time the clip occurs in the multimedia document. The UI uses this information to construct the query response pages displayed to the user.

The librarian is built on an Oracle relational database running on Tru64 Unix 4.0. The use of a central repository for shared information allows a robust distributed architecture which can scale on demand.

2.4. Indexer

The indexer provides an efficient catalogue of audio and video documents based on the transcription produced by the speech decoder. As well as supplying the user interface with a list of documents that match a user's query, the indexer also retrieves the location of these matches within the documents. It does this using a modified version of the AltaVista query engine.

The indexer sorts the matches according to relevance, as described in Section 3.3. We define relevance using the term frequency inverse document frequency (tf/idf) metric [7], adjusted for the proximity of the terms within the document. This sorting is performed on both the list of documents returned and the list of match locations. In addition to the transcription of the audio, we also index the meta-data if any is available.

2.5. User Interface

The Web server passes the user queries to the indexer, reads back the list of matches, retrieves the associated meta-data from the librarian, and formats and displays the results. As part of this process, it also performs the advanced functions described in Section 4, such as highlighting the matches within the transcript and expanding and normalising acronyms, abbreviations and numbers.

The Web server returns up to 20 pages of up to 10 documents each sorted by relevance. External links to the original audio and video files are displayed, as well as a link for each document to a page with more details. These details include a navigable timeline of the 20 most relevant matches within the document, and also an option to display between 30 seconds and 2 minutes of the transcript text surrounding the match.

3. PERFORMANCE ANALYSIS

In this section we obtain objective measurements of each of the components of the *SpeechBot* system. We describe the performance of the transcoder, the speech recognition engine and the information retrieval engine (a combination of the speech recognition output and the indexer).

3.1. Transcoder

The current version of the transcoders can handle a variety of different formats both streaming and not streaming. 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 downloadable hours of audio per day.

3.2. Speech Recognition

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 and Rick Emerson among others. 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 6.5 Kbps RealAudio codec. 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, etc. The selected segments are transcribed manually, and the transcription is used to estimate the Word Error Rate ($WER = (S+I+D)/T$) where S, I, D and T are the number of substitutions, insertions, deletions and total number of words respectively.

We use a language model trained using the DARPA broadcast news Hub-4 1998 text corpora [5]. It contains a vocabulary of 64,000 words which corresponds to 4 million bigrams and 15 million trigrams in the language model.

We explored two acoustic modeling approaches. In our first approach we build acoustic models by training 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 had to be up-sampled to 16 kHz. The second approach used models trained on the same training corpus but after being encoded using the 6.5 Kbps RealAudio codec, and then decoded to a sampling rate of 8 kHz. This encoding/decoding operation was performed to reduce the acoustic mismatch between the training corpora and the testing corpora.

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 RealAudio encoded/decoded models were used. These results are for a system with 16 Gaussian mixture components per HMM state and 6000 shared states. Table 1 presents full results for the 8 kHz RealAudio system.

Show Name	Chunk 1	Chunk 2	Chunk 3	Chunk 4
Art Bell	50.5%	46.3%	51.2%	44.5%
Dr. Laura	51.2%	47.4%	52.7%	59.2%
R. Emerson	48.2%	51.5%	53.1%	56.4%

Table 1: Speech Recognition WER. 8 kHz RealAudio modes, 16 Gaussian mixture components per clustered HMM state. 6000 cluster states.

3.3. Information Retrieval Performance

The retrieval performance was measured with independent and unbiased testers. They evaluated a set of queries selected from the list of 100 most frequently submitted queries. Each tester performed 40 queries, and assessed the precision of the top shows returned for each. The study was limited to the top 20, 10 and 5 shows, based on the assumption that typical users tend only to look at the first couple of pages of the retrieved results [8]. 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 and proper nouns. Example queries are *bill clinton*, *internet* and *Y2K*.

Testers judged a document as relevant if and only if the concept described in the query was spoken in the retrieved document. To assess whether a given result was relevant, testers read the transcripts returned and listened to the corresponding parts of the show. To evaluate the retrieval results, we used a standard average precision metric [7]. We did not run the experiment on a standard test set, such as TREC-SDR [3], but rather on real data (both audio content, and queries). We believe this approach makes this experiment unique.

Table 2 presents average retrieval precision numbers for the above described experiments for two different ranking functions.

Rank function	docs (count)	time (hours)	queries (count)	Prec.
A	3111	3000	40	69%
B	5081	4862	40	85%

Table 2: Average retrieval precision for whole documents using two different ranking functions.

Ranking function A scores documents using a term frequency inverse document frequency metric [7], combined with scores based on the proximity of query terms and their location within the document (the closer to the beginning of the document a term, the higher its score). The proximity bias helps to retrieve documents with a multi-word

query string. Ranking function B scores documents using the same term frequency inverse document frequency metric and proximity bias, but the location based score is replaced by a query term frequency score within the document (the more frequent the occurrence of the term is, the higher its score).

The retrieval performance of the system is better than expected considering the accuracy of the speech decoder, and we postulate two reasons for this: 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.

Retrieval errors were due to two main reasons. First, insertion or substitution recognition errors cause query words to appear erroneously in the transcripts. The ranking function A is particularly sensitive to this kind of error. We observed several cases where an insertion at the very beginning caused the document to get erroneously a high score. This kind of false alarm error represents roughly half of the cases of the appearance of non-relevant documents. The ranking function B helps to alleviate this problem, and explains partially the improvement observed. The second main reason for a non-relevant document to be retrieved is when the query words are mentioned out-of-context, or when they are inherently ambiguous. For example, the query *AIDS* returned many documents which talked about *aids* meaning *helps* rather than a disease.

4. USABILITY STUDIES

Although the interface is very similar to most text-based search engines, users encountered several difficulties. When performing multiword searches, in general users expected to see all query words in the results. Interestingly, even regular *AltaVista* users did not understand the query syntax. For instance, when looking for *divorce in Dr. Laura's interviews*, most first tried *laura divorce*, clever ones then tried *laura+divorce*, but nobody tried *+laura +divorce*. This problem was addressed by exposing the possible relationships between search words *(all/any/phrase/Boolean)* in a drop down menu and making *all* the default.

Second, the syntax for query terms is sometimes non-intuitive. For example, users may type acronyms in different forms: *USA* or *U.S.A.*, but without special processing not all written forms will match their spoken counterpart. To alleviate this problem, terms like abbreviations, acronyms, and numbers are now expanded and normalized by the system to a common form for both querying and indexing. For instance, the number 250 will be trained and recognized as *two hundred fifty*, likely be searched as 250 and will be indexed in its normalized form as 250.

5. CONCLUSIONS 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 both with and without transcriptions. The current version of the system is capable of indexing up to 800 hours of audio per 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. Our experiments show acceptable retrieval accuracy despite high recognition error rates. This suggests that indexing of audio documents on the Web is feasible given the 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 alternative ways of indexing the output of the speech recognizer. Finally, 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.

6. ACKNOWLEDGMENTS

We first would like to thank Andrew Leonard, for his world class web master support. He has been in charge of running the site and keeping the index fresh. We also would like to aknowledge the past contributions of Mike Swain, Dave Goddeau, Anna Litvinova, R. Paul Johnson and Chris Weikart. Finally we would like to thank Frank Bomba, Ron Gentile, Gene Preble, and Ben Jones for their support for testing and evaluating the system.

7. REFERENCES

[1] H. D. Wactlar, A. G. Hauptmann, and M. J. Witbrock, "Informedia: News-on-demand experiments in speech recognition," in *DARPA Speech Recognition Workshop*, 1996.

[2] S. E. Johnson, P. Jourlin, G. L. Moore, K. S. Jones, and P. C. Woodland, "The cambridge university spoken document retrieval system," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1999.

[3] D. Abberley, G. Cook, S. Renals, and T. Robinson, "Retrieval of broadcast news documents with the thisl system," in *Proceedings of the 8th Text Retrieval Conference (TREC-8)*, 1999.

[4] M. Witbrock, "Search engines and beyond," in *SearchEngine Meeting, Boston*, 1999.

[5] D. S. Pallet, J. G. Fiscus, J. S. Garafolo, A. Martin, and M. Przybocki, "1998 broadcast news benchmark test results," in *DARPA Speech Recognition Workshop*, 1999.

[6] P. J. Moreno, C. Joerg, J. M. Van-Thong, and O. Glickman, "A recursive algorithm for the forced alignment of very long audio segments," in *International Conference on Spoken Language Processing*, 1998.

[7] G. Salton and M. J. McGill in *Introduction to Modern Information Retrieval. McGraw-Hill*, 1983.

[8] C. Silverstein, M. Henzinger, H. Marais, and M. Moricz, "Analysis of a very large altavista query log," in *SRC Technical Note 1998-014*, 1998.