ISCA Archive
http://www.isca-speech.org/archive

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

AN EXPERIMENTAL STUDY OF AN AUDIO INDEXING SYSTEM FOR THE WEB

Beth Logan *Pedro Moreno* *Jean-Manuel Van Thong* *Ed Whittaker*

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

ABSTRACT

We have developed a speech recognition based audio search engine for indexing spoken documents found on the World Wide Web. Our site (http://www.compaq.com/speechbot) indexes around 20 news and talk radio shows covering a wide range of topics, speaking styles and acoustic conditions from a selection of public Web sites with multimedia archives. In this paper, we describe our system and its performance, focusing on the speech recognition and retrieval aspects. We describe our training procedure in some detail and report our historical error rate since the site launch. We also investigate the impact of Out Of Vocabulary (OOV) words. Finally we report the results of retrieval experiments which demonstrate that our system can index effectively.

1. INTRODUCTION

In recent years, much research in the speech recognition community has focused on broadcast news tasks. This has led to many advances but now is the time to move away from such constrained studies and look at more real-world applications. At Cambridge Research Laboratory, we have developed an audio search engine which indexes spoken documents found on the World Wide Web [11]. Our current site indexes a number of news and talk radio shows from a selection of public Web sites with multimedia archives. Speech recognition technology is the key to indexing these shows as in most cases a transcription is not available.

Compared to the more 'sanitized' data used in the Broadcast News evaluations, found audio on the web is both acoustically and linguistically more challenging to transcribe. The radio shows we index cover a wide range of topics and speaking styles. The audio from these shows suffers in acoustic quality due to bandwidth limitations, coding, compression, and poor acoustic conditions. Regardless, we show we can achieve accuracy satisfactory for indexing audio from the Web.

A number of other groups have also built indexing systems based on speech recognition (e.g. [12], [3], [5], [1]). We differ from these projects in several ways. First, we fetch our audio documents directly from the Web. Second, we do not retain the original content but rather keep only a link to the indexed document, similar to traditional search engines. Finally, our system is designed to scale on demand.

The purpose of this paper is to describe our system and its performance, focusing in particular on the speech recognition and retrieval aspects. We begin by describing the main components of our indexing system. Next, we describe the speech recognition module in some detail and present the historical word error rates since the system launch. We also investigate the impact of the static dictionary and language model we are currently using. Finally, we report the results of retrieval experiments, giving an ultimate performance metric for our system, and present suggestions for future work

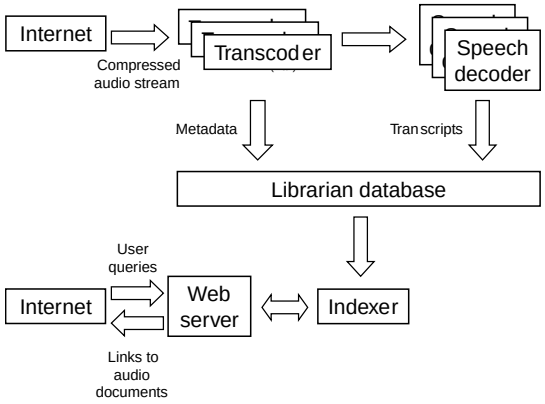

Figure 1: Overall architecture of the system

2. THE SPEECHBOT SYSTEM

SpeechBot (www.compaq.com/speechbot) [11] is a public search system similar to AltaVista which indexes audio from public Web sites such as Broadcast.com, NPR.org, Pseudo.com and Internet-News.com. The index is updated daily as the audio of new shows is archived on the Web. As of June 2000, our index serves about 5000 hours of audio, growing at the rate of about 100 additional hours per week. We do not retain this content however for both copyright and storage space reasons. Thus our Website is similar in spirit to other Web search sites as it contains an index rather than the actual multimedia content.

Our indexing system consists of the following modules: the transcoders, the speech decoders, the librarian database, and the indexer as shown in Figure 1. We briefly describe each of these components in the following sections.

2.1. Transcoders

The transcoders fetch and decode video and audio files from the Internet. For each document, they obtain both meta-data (such as the sample rate and document title) and the raw audio. Typically the raw audio is RealAudio compressed at 6.5 kbps. This audio is downloaded to a temporary local repository and converted to 8kHz PCM wav format.

2.2. Speech Decoders

For speech recognition, we use Compaq's Calista system. This is a large vocabulary continuous speech recognition package which uses standard Hidden Markov Model (HMM) technology. Calista yields an error rate of about 20% on a single pass search on the 1998 ARPA HUB4 evaluation corpora [7] with an average computational load of 6 times real time on a Compaq DS20 EV6 Alpha processor running at 500 MHz. The production system consists of a farm of 30 Compaq dual Pentium II/IIIs (450 MHz and 600 MHz) with 256 Mb RAM running Linux 6.0. We shall discuss the speech recognition system in more detail in Section 3. When the transcription is available, we can replace the speech recognition module with an aligner module.

2.3. Librarian

The librarian manages the workflow of the tasks required to index documents. Each registered process of the workflow can send a request to the librarian for work to be performed. This includes tasks such as speech decoding, text/audio alignment or insertion into the index. This centralized model leads to a robust distributed architecture which can scale on demand.

2.4. Indexer

The indexer module catalogs the documents based on the transcription produced by the speech decoder using a modified version of the AltaVista query engine [2]. 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.

Our ranking algorithm scores documents using a term frequency inverse document frequency metric [8], combined with scores based on the proximity of query terms. The proximity bias helps to retrieve documents with a multi-word query string.

3. SPEECH RECOGNITION PERFORMANCE

In this section we describe the training procedure for our speech recognition system and investigate its performance. We see that the error rates obtained fall far short of the standard Broadcast News benchmarks due to the adverse nature of the audio. However, the primary metric for judging the performance of our indexing system is information retrieval results. As noted by others (e.g. [4]) and demonstrated in Section 4, it is still possible to obtain good retrieval performance despite poor error rates. In addition, as noted in Section 4, even if perfect transcriptions could be obtained, errors in retrieval due to for example word ambiguities would still occur. Nevertheless, high error rates are still undesirable since they do impact retrieval performance. Also better recognition rates would improve the quality of our user interface which displays the recognized transcripts.

3.1. System Details

Calista is a standard HMM-based speech recognizer. It uses 3-emitting-state Gaussian mixture HMMs to model triphones. The models are trained in the following (very standard) manner. First 49 context-independent phonemes with 1 Gaussian mixture component per state are trained, starting from 'flat' (i.e. equivalent) models. These context independent models are then cloned to give triphone models which are pruned so the total number of states is 6000. Starting from these 6000-state models, successive iterations of Baum-Welch training and mixture splitting result in models containing up to 16 Gaussian mixture components per state.

We train our acoustic models on MFCC cepstral coefficients augmented with delta and acceleration coefficients generated from around 100 hours of the 1997 and 1998 Broadcast News corpus provided by LDC [10]. To more closely match the audio to be recognized, we use a modified version of the training corpus which has been encoded using the Real Audio encoder and decoded to to a sampling rate of 8kHz. We have previously shown that the use of these Real Audio acoustic models decreases the absolute error rate of a 5 hour test set by around 10% absolute from 60.5% to 49.6% for a 16 mixture component Gaussian system [11].

In conjunction with these acoustic models, we use a standard trigram language model trained using the DARPA broadcast news HUB4 1997 and 1998 text corpora augmented with additional text from News.com. It contains a vocabulary of 64000 words, corresponding to 4 million bigrams and 15 million trigrams.

For speed reasons, we run a single pass decoder with no additional adaptation and 8 mixture components per Gaussian. Experiments detailing our speed/accuracy tradeoffs are described in [11]. We use a very simple but robust segmentation technique, breaking the audio into 35s segments which are overlapped by 5s. We then reconstruct the transcripts by joining segments together at the overlap.

3.2. Recognition Performance Over Time

Since November 1999, we have been monitoring the error rate of our system by collecting and transcribing an hour's worth of content at 2-3 week intervals. These test sets each consist of 4x15 minute segments randomly selected from approximately one week's worth of content (according to how much audio was available due to disk limitations). We take this 'sampling' approach to monitoring the performance of our system since resource restrictions prevent us from transcribing all 5000 hours of our content.

Figure 2 shows the error rate for the series of 15 test sets tested by our production system (i.e. a one pass decoder with 8 Gaussian mixture components per state). We see that while there have been some fluctuations, the error rate has remained around the

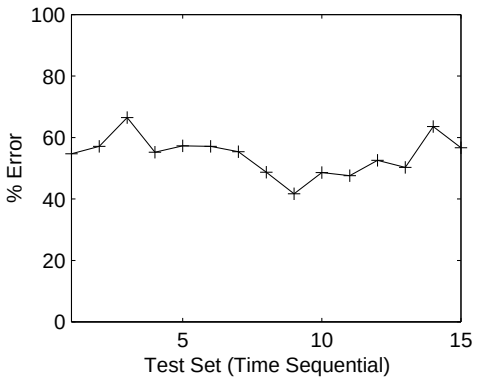

Figure 2: Error rate of 15 1 hour test sets collected at 2-3 week intervals from November 1999 to June 2000

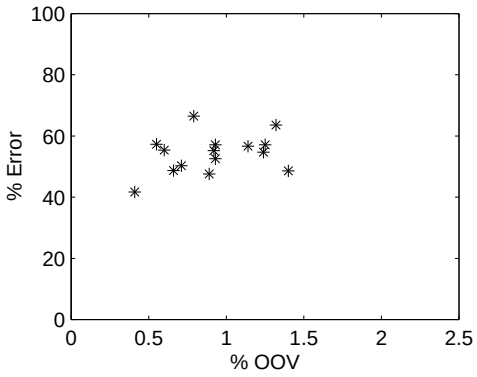

Figure 3: Error rate of 15 test sets as a function of OOV rate

55% mark since the system launch. The fluctuations are due in part to the high variability and large amount of the content *vs* the comparatively small size of the test sets. The error rate is clearly much higher than typical Broadcast News error rates (around 20%). However, as we shall see in Section 4 this performance is sufficient for indexing.

3.3. Impact of the Static Dictionary

To date, Speechbot has been running with a 64000 word dictionary and language model built in late 1999. However, many of the shows we index are topical news programs. Over time as new topics replace old, we expect the vocabulary to change.

Figure 3 shows the word error rate of the 15 test sets *vs* the Out Of Vocabulary (OOV) rate. We see from this figure that our OOV rate is very low and only loosely correlated with the error rate. To investigate the effect of OOV errors, we conducted an artificial experiment where we found the OOV words for several test sets and inserted these into the vocabulary. We then rebuilt the language model and dictionary and reran the recognition experiment for each set. Although around half the OOV words were not found in the language model training text and therefore had very low language model probabilities, this experiment still gives a feel for the improvement we could expect if appropriate texts containing the OOV words were available.

Table 1 shows the result of this experiment. We see that since our OOV rate is so low, improving it will have only a minor impact on our recognition performance. Therefore it is likely that other factors such as acoustic mismatch, pronunciation inaccuracies and language model mismatch are the main contributors to our error rate. Future work will therefore focus on these areas.

Test Set	OOV Rate	Error Rate
12-3-1999	1.3%	57.1%
	0.0%	56.7%
2-24-2000	1.4%	48.6%
	0.0%	48.4%
5-2-2000	1.2%	63.6%
	0.0%	63.5%

Table 1: Impact of reducing OOV rates for the test sets with the worst OOV rates.

4. RETRIEVAL PERFORMANCE

In this section, we describe the results of information retrieval experiments and additionally investigate the number of OOV words in user queries.

We evaluated a set of 50 queries on an index of 4188 hours of content, representing 4695 programs. The queries were selected from the list of the 100 most frequently submitted queries to the public site since its launch. The words were selected such that they cover a large variety of topics, varying length of words (phonemewise), and varying types of words such as acronyms and proper nouns. None of the queries selected were OOV.

The average retrieval precision results are shown in Table 2. We report the standard retrieval precision P given by $P = N/T$ where N is the number of relevant documents and T is the total number of documents retrieved [8]. We report results for $T = 5,\ 10 \text{ and } 20$. That is, we only consider the first 5, 10 and 20 returned shows since users tend to only look at the first couple of pages of retrieved results [9].

Number of Documents	Average Precision
5	87.8%
10	83.9%
20	77.5%

Table 2: Average retrieval precision for the top 5, 10 and 20 documents.

From this table we see that our system has usable performance. Examination of the non-relevant documents returned showed that errors were due to two main reasons. First, insertion or substitution recognition errors cause query words to appear erroneously in the transcripts. This was the cause of about half the errors. The second main cause of error 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.

The retrieval performance of our system is better than expected considering the accuracy of the speech decoder. We believe this is for several reasons. First, the query words are often repeated several times during a show and are thus more likely to be recognized. Second, the speech recognizer tends to make fewer mistakes on keywords since these are on average longer.

4.1. OOV Rates of Queries

The above experiment is somewhat artificial since none of the query words were OOV. In fact, as shown in Table 3, a large proportion of the query words *are* OOV. This table shows results for all queries received since the launch of the site. The first line of this table shows OOV results for the dictionary used in our production system. We show both the average percentage of OOV words per query and the average OOV rate overall.

Vocabulary derived from	Average OOV rate per Query	Weighted OOV
Broadcast news dictionary	16.0%	12.6%
As above + transcript words	15.8%	12.4%

Table 3: Average OOV Rate per Query and Weighted OOV for various vocabularies

These OOV query words fall into several categories:

- company and proper names
- words with text normalization errors (mostly acronyms)
- foreign or misspelt words
- words using wildcards or other unsupported query syntax
- other words (e.g. rare words, rude words)

While we could improve our text normalization and refine our user interface to disallow unsupported query syntax and perhaps catch misspellings, increasing the vocabulary to include the other OOV words is more challenging. The second line of Table 3 shows OOV percentages if the vocabulary is increased to include all new words in the 15 hours of transcripts we have available. We see this has only a minor impact on the percentages since the amount of transcribed data available is comparatively small. One issue raised is how many of the OOV words are actually relevant to our index. Without ground-truth transcripts this is impossible to answer. However, we believe that by refining our index to include categories of words we may be able to provide our users with a richer experience. Another approach is to investigate subword based retrieval which does not need a dictionary (e.g. [6]). The use of these methods is the subject of ongoing work.

5. CONCLUSIONS

In this paper we have described a novel speech recognition based audio indexing system which enables users to search spoken documents found on the World Wide Web. We have described a new methodology based on sampling to *measure* the performance of our system when the size of the index is extremely large, more than 5000 hours of audio and video content, and when the ground truth is not readily available. We have shown that although our word recognition error rates are fairly high (around 55%) we are still able to create a highly usable index.

A major portion of this paper has investigated the effect of OOV words on the system performance. Surprisingly, we found that OOV words have little impact on the error rate of the speech recognizer. We have also found that contrary to common belief, the OOV rate has not changed greatly over the 8 months period in which the site has been active.

We have also investigated the OOV of user queries. We have found this OOV (12%) to be considerably higher than the speech recognition OOV. Clearly more work is needed to investigate the implications of this high rate. This rate could be improved by altering our text normalization and user interface. Future work will also focus on minimizing the effect of OOV query words by investigating indexing word categories and/or performing subword based retrieval. The latter should also alleviate the problem of OOV words in the speech recognition module.

6. REFERENCES

1. 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.
2. M. Burrows. *Method for Indexing Information of a Database.* U.S. Patent 5,745,899, 1998.
3. J. Garfolo, E. Vorhees, C. Auzanne, V. Stanford, and B. Lund. Spoken document retrieval track overview and results. In *Proceedings of the 7th Text Retrieval Conference (TREC-7)*, 1998.
4. A. Hauptmann, R. Jones, K. Seymore, S. Slattery, M. Witbrock, and M. Siegler. Experiments in information retrieval from spoken documents. In *Broadcast News Transcription and Understanding Worshop*, pages 175–181, 1998.
5. 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.
6. K. Ng and V. Zue. Subword unit representations for spoken document retrieval. In *European Conference on Speech Communication and Technology*, 1997.
7. D. S. Pallet, J. G. Fiscus, J. S. Garafolo, A. Martin, and M. Przybocki. 1998 brodcast news benchmark test results. In *DARPA Speech Recognition Workshop*, 1999.
8. G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.
9. C. Silverstein, M. Henzinger, H. Marais, and M. Moricz. Analysis of a very large AltaVista query log. Technical report, SRC, October 1998.
10. R. M. Stern. Specification of the 1996 Hub 4 broadcast news evaluation. In *DARPA Speech Recognition Workshop*, 1997.
11. J-M. Van Thong, D. Goddeau, A. Litvinova, B. Logan, P. Moreno, and M. Swain. Speechbot: a speech recognition based audio indexing system for the web. In *Proc. International Conference on Computer-Assisted Information Retrieval (RIAO)*, 2000.
12. H. D. Wactlar, A. G. Hauptmann, and M. J. Witbrock. Informedia: News-on-demand experiments in speech recognition. In *DARPA Speech Recognition Workshop*, 1996.