# THE CAMBRIDGE UNIVERSITY SPOKEN DOCUMENT RETRIEVAL SYSTEM

*S.E. Johnson*†*, P. Jourlin*‡*, G.L. Moore*†*, K. Spärck Jones*‡ *& P.C. Woodland*†

†Cambridge University Engineering Department,Trumpington Street, Cambridge CB2 1PZ, UK.
Email: {sej28,glm20,pcw}@eng.cam.ac.uk
‡Cambridge University Computer Laboratory, Pembroke Street, Cambridge, CB2 3QG, UK.
Email: {pj207,ksj}@cl.cam.ac.uk

## ABSTRACT

This paper describes the spoken document retrieval system that we have been developing and assesses its performance using automatic transcriptions of about 50 hours of broadcast news data. The recognition engine is based on the HTK broadcast news transcription system and the retrieval engine is based on the techniques developed at City University. The retrieval performance over a wide range of speech transcription error rates is presented and a number of recognition error metrics that more accurately reflect the impact of transcription errors on retrieval accuracy are defined and computed. The results demonstrate the importance of high accuracy automatic transcription. The final system is currently being evaluated on the 1998 TREC-7 spoken document retrieval task.

## 1. INTRODUCTION

There is now widespread use of information retrieval (IR) techniques to access information stored in electronic texts. One of the most widely used examples of this is in internet search engines. However, there is also much information contained in "documents" that are not initially created as text but are spoken. One such area is the audio associated with radio and television news broadcasts. If these audio sources could be transcribed automatically then the information they contain can be indexed and relevant portions of broadcasts retrieved using conventional IR techniques.

Recently there has been increasing interest in this spoken document retrieval (SDR) task and the first international evaluation of spoken document retrieval techniques took place in 1997 as part of the TREC-6 (Text REtrieval Conference) [2]. This involved using a rather simple "known item" paradigm in which it is known in advance that only a single document is relevant to any given query. For the TREC-7 SDR track more general and more difficult query types are being investigated and evaluated using the standard IR measures of precision and recall.

This paper describes the experiments we have done to develop an SDR system for the TREC-7 evaluation. As part of this effort, we have automatically transcribed the TREC-6 test data using two different HTK-based transcription systems and evaluated the IR performance using a set of 60 queries (denoted CU60) and corresponding relevance assessments that we developed for this purpose. Of particular interest is the relationship between transcription accuracy and retrieval performance and we have investigated this relationship using the two HTK-based recognisers as well as transcriptions from a baseline IBM recogniser and those supplied by Sheffield University. These four sets of automatic transcriptions provide a wide range of word error rates.

The paper is organised as follows. Firstly an overview of the HTK broadcast news transcription system used to generate high quality automatic transcriptions is given. Then we give an overview of the information retrieval problem and the approach we have adopted. A number of transcription error metrics are then introduced and the relationship between them presented.

Finally we describe improvements to the system that we developed for the TREC-7 evaluation which required the transcription of nominally 100 hours of broadcast audio.

## 2. OVERVIEW OF THE HTK BROADCAST NEWS TRANSCRIPTION SYSTEM

The input data is presented to the HTK transcription system as complete episodes of broadcast news shows and these are first converted to a set of segments for further processing. The segmentation uses Gaussian mixture models to divide the audio into narrow and wide-band audio and also to discard parts of the audio stream that contains no speech (typically pure music). The output of a phone recogniser is used to determine the final segments which are intended to be acoustically homogeneous.

Each frame of input speech to be transcribed is represented by a 39 dimensional feature vector that consists of 13 (including $c_0$) cepstral parameters and their first and second differentials. Cepstral mean normalisation (CMN) is applied over a segment.

The system uses the LIMSI 1993 WSJ pronunciation dictionary augmented by pronunciations from a TTS system and hand generated corrections. Cross-word context dependent decision tree state clustered mixture Gaussian HMMs are used with a 65k word vocabulary. The full HTK system [5] operates in multiple passes and incorporates unsupervised maximum likelihood linear regression (MLLR) based adaptation and uses complex language models via lattice rescoring and quinphone HMMs. This system gave a word error rate of 16.2% in the 1997 DARPA Hub4 broadcast news evaluation.

For the experiments on the TREC-6 test data triphone HMMs were trained on nominally 50 hours (actually about 35 hours) of broadcast news audio and the language models trained on 132 million words of broadcast news texts, the LDC-distributed 1995 newswire texts, and the transcriptions of the acoustic training data.

Two versions of the HTK system were used in the experiments on TREC-6 SDR test data. The first is essentially the first pass of the system in [5] and runs in about 45 times real time on a Sun Ultra2. This HTK-1 system uses gender independent (but bandwidth dependent) HMMs and a trigram language model. The HTK-1 system gave a word error rate of 28.6% on the 50 hours (1451 stories/documents) of the TREC-6 SDR test data.

The second HTK system (HTK-2) uses the same gender independent HMMs as HTK-1 in a faster first pass. The output of the first pass is used along with a top-down covariance-based segment clustering algorithm to group segments within each show to perform unsupervised test-set adaptation using MLLR. A second recognition pass using a bigram language model was used to generate word lattices using adapted gender and bandwidth specific HMMs. These bigram lattices were expanded using a 4-gram language model and the best pass through these lattices gives the final output. This system gave a word error rate of 24.1% on the TREC-6 test data and ran in about 50 times real time. A similar system to this (but with the full Hub4 training data) gave a 17.6% word error rate on the 1997 Hub4 evaluation test data.

We have also used alternative automatic transcriptions to assess the effect of error rate on retrieval performance. NIST supplied a baseline transcription (computed by IBM) with 50.0% word error rate (WER) and the transcription obtained by Sheffield University [1] had a 39.8% WER.

## 3. INFORMATION RETRIEVAL SYSTEM

### 3.1. Background

In an information retrieval system the user generates a request or *query* for the information they would like to obtain. This may take the form of a command (e.g. "Find me information about El Nino") or a question ("Have there been any volcanic eruptions in Montserrat recently?"). The information retrieval engine then presents the user with a ranked list of documents, ideally with those most relevant to the query topic first, from which the user can obtain the information, or answers to questions, they are seeking. [1]

In order to be able to score the match between the documents and the query, a method of representing, (i.e. *indexing*) the linguistic information in both the documents and the query must be found along with a way of matching the two. Since the information in the documents is encapsulated in their words, these offer a starting point for the representation of a document. Words which are thought to contain no particular information, are normally removed from documents in a process called *stopping*. These *stopwords* include function words, such as "a" and "the".

Furthermore the suffixes on the words are removed using a stemming algorithm. This extracts the information-rich part of the word for example converting "managing", "manage", "manageable" and "managed" all to "manag". This allows several different linguistic methods of conveying the same information to be treated equally.

Finally an index file is created which contains the number of times a word occurs in each document and the number of documents each word occurs in. This then acts as the input (along with the stemmed/stopped query) to the main retrieval engine.

The success of an IR system is measured in terms of *precision* (the proportion of the returned documents which are relevant) and *recall* (the proportion of the relevant documents which are returned). High precision means few irrelevant documents are presented to the user whereas high recall means the system has found most of the relevant documents.

[1] The more difficult task of directly answering a question posed in the query is not considered to lie within the field of information retrieval

### 3.2. IR System Details

The system uses a stop word list and performs stemming with the widely-used Porter stemming algorithm [3]. Unfortunately the Porter algorithm suffers from some problems for example mapping "news" to "new" but not equating "government" and "governmental". To overcome these deficiencies a stemmer exceptions list is used to map known problem-words to their stems directly.

For ease of implementation, this stemming exceptions list is incorporated into a mapping stage which standardises spellings for commonly mispelt words or phrases such as "Chechnia/Chechnya" or "all right/alright" and makes synonyms such as "United States" and "U.S." equivalent. After this stage each document is represented by a stopped/stemmed/mapped unordered list of words. The effect of these processes on the number of words at each stage using the IR system for the different automatically generated transcriptions is shown in Table 1.

| Recogniser | original | +stop | +map | +stem* |
|---|---|---|---|---|
| IBM Baseline | 404559 | 188117 | 187595 | 184214 |
| Sheffield | 382855 | 186447 | 185937 | 182864 |
| HTK-1 | 397942 | 183475 | 182869 | 178805 |
| HTK-2 | 393592 | 185527 | 184951 | 181105 |

* This also includes dealing with abbreviations

Table 1: Number of Words for TREC-6 SDR after various stages of processing

Whilst the simplest way to generate a matching score for a document to a query is to count the number of shared stopped/stemmed terms, this is far too crude and modern systems use term *weighting*. This exploits information about the document length, $dl(j)$, the number of documents containing the query term, $n(i)$, and the number of times the term occurs in the given document, $tf(i,j)$.

Specifically, our system generates a *combined weight*, $cw(i,j)$ for document $j$ and matched query term $i$ in the same way as the systems developed by City University[4] so that

$$cw(i,j) = \frac{(\log N - \log n(i)).tf(i,j).(K+1)}{K.(1-b+b.ndl(j)) + tf(i,j)}$$

where $N$ is the number of documents in the collection, $ndl(j)$ is the length of document $j$ normalised by the average $dl$ and $K$ and $b$ are tuning constants.

| | Stop1+Stem | Stop2+Stem | Stop2+Stem+Map |
|---|---|---|---|
| manual | 0.6687 | 0.6758 | 0.6960 |
| HTK-2 | 0.6287 | 0.6478 | 0.6746 |

Table 2: Average precision for the CU60 queries using different stop lists and mapping file

We have developed the IR system for the broadcast news task and improved the stop list and the map file list for this task. To show the effect of these changes Table 2 gives the average precision of the system using the HTK-2 and the manual transcriptions for the TREC-6 test data with the CU60 queries. It can be seen that the revised stop-word list (stop2 vs stop1) and the mapping file considerably improve the average precision measure.

## 4. TRANSCRIPTION ERROR METRICS FOR IR

Speech recognition accuracy is conventionally expressed in terms of word error rate (WER). To calculate this an alignment of the hypothesised and reference transcriptions is made and the number of insertion $(I)$ , deletion $(D)$ and substitution $(S)$ errors are found. For $W$ words in the reference transcription, the word error is then given by:

$$\text{WER} = \frac{(S+I+D)}{W}.100\%$$

This is a reasonable error measure for pure transcription tasks but because of the way that IR systems manipulate the transcription before use (stopping, mapping, stemming and ignoring word order), WER may not be the most appropriate error measure for an IR system.

| Recogniser | WER | +stop | +map | +stem |
|---|---|---|---|---|
| IBM Baseline | 50.0 | 47.5 | 47.2 | 44.3 |
| Sheffield | 39.8 | 37.6 | 37.1 | 34.6 |
| HTK-1 | 28.6 | 24.9 | 24.7 | 22.2 |
| HTK-2 | 24.1 | 21.5 | 21.2 | 18.7 |

Table 3: Story-based % Word Error Rates for TREC-6 data

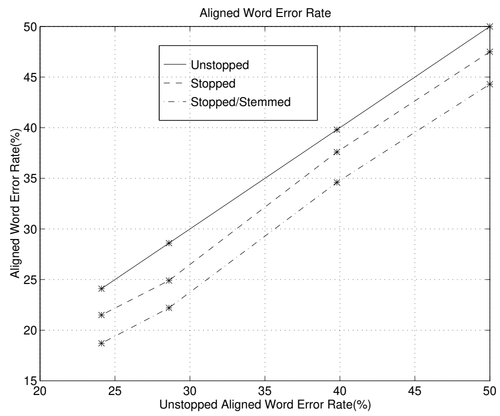


Figure 1: Correlation between Word Error Rates with and without stopping, mapping and stemming

The WER before and after stopping, with mapping and stemming are given in Table 3 and the relationship between them and the unstopped WER is shown in Figure 1. These results show that stopping and stemming the documents reduces the WER whilst it can be seen from Table 1 that the number of words in the document collection is also reduced. It is interesting to note that the unstopped WER offers a good predictor of both stopped and stemmed WER.

We have investigated several alternatives ways of measuring the effective number of errors going into the main retrieval system and compared them to the traditional WER. These account for the effects of stopping and stemming and also the lack of importance of word order. For such purposes, conventional substitution errors could be regarded as two errors since not only do they miss an occurrence of a word, they also add a spurious one.

We first define a new error rate metric called a *Term Error Rate (TER)* which is independent of word order and given by:

$$\text{TER} = \frac{\sum_w |A(w) - B(w)|}{W}.100\%$$

where $A(w)$ and $B(w)$ represent the number of times word $w$ occurs in the reference $A$ and the transcription $B$. This also models a traditional substitution error as two errors.

| Recogniser | Quasi-TER | TER | +stop | +stem |
|---|---|---|---|---|
| IBM Baseline | 83.6 | 61.1 | 73.7 | 67.4 |
| Sheffield | 65.1 | 48.4 | 59.2 | 53.0 |
| HTK-1 | 46.1 | 32.9 | 37.6 | 32.8 |
| HTK-2 | 38.7 | 28.2 | 32.8 | 28.2 |

Table 4: % Term Error Rates for TREC-6 data computed by story (document)

The TER for unstopped, stopped and stemmed document sets is given in Table 4 along with a Quasi-TER for the unstopped case which is predicted from the transcription alignment by

$$\text{Quasi-TER} = \frac{(2S+I+D)}{W}.100\%$$

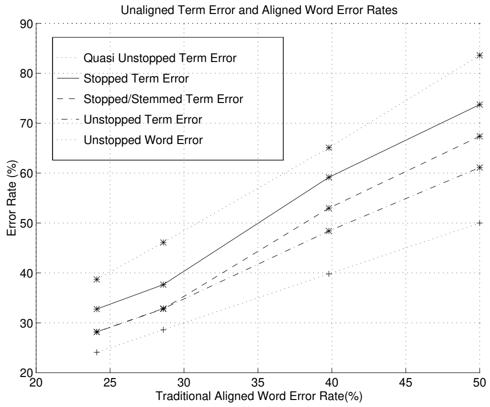


Figure 2: Relation between WER and TER

The results given in Table 4 allow the separate influences of counting substitutions twice and ignoring the word order to be seen. The relationship between the TER and WER is shown in Figure 2. It is interesting to note that there is a large difference between the Quasi- and True- unstopped TER, suggesting many of the errors occurring with alignment cancel out. Furthermore, stopping the documents increases term error rate, although it was previously shown to decrease (aligned) word error rate. This is thought to be because the majority of cancelling errors occur with the shorter, stopped words, so the cancelling effect is reduced by stopping, hence increasing TER. Stemming will always reduce both WER and TER.

## 5. SDR SYSTEM RESULTS

The precision-recall results for the CU60 queries, using the IR engine on the HTK-1, HTK-2, IBM baseline and Sheffield transcriptions, along with the manual reference transcriptions, are shown in Figure 3. A graph of average precision against unstopped aligned word error rate and stopped/stemmed term error rate is given in Figure 4.

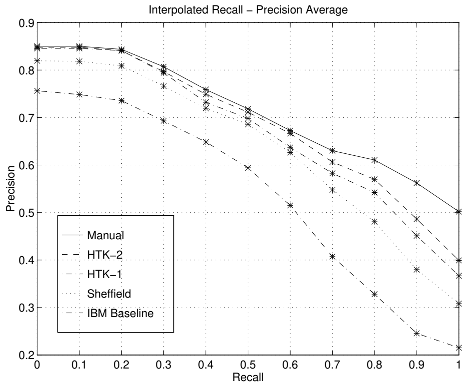


Figure 3: Precision-Recall graphs for varying WER

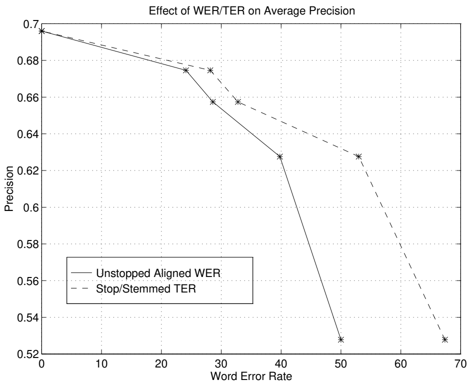


Figure 4: Effect of word error rate on Average Precision

It can be seen that IR performance decreases as both unstopped aligned WER and stop/stemmed TER increase. It is also interesting to note that as WER increases above about 30% the retrieval performance starts to fall off rapidly. It is pleasing to see that at low recall levels (where the number of documents returned to the user is relatively small), the precision is similar for both the manual transcription and the two versions of the HTK output. The graphs also show that the average precision is affected by recognition errors and that high accuracy transcription is required for the average precision to be close to that of the manual transcription.

## 6. TREC-7 SDR SYSTEM

For the 1998 TREC-7 SDR task about 100 hours of broadcast news had to be transcribed but the full 1997 Hub4 training set was available for HMM estimation. We therefore used a new set of HTK triphone HMMs in an HTK-2 recognition setup. Furthermore new language models were estimated from an extended set of broadcast news transcripts and newspaper texts that covered the epoch of the test data. The 65k vocabulary was also tuned. We estimate that the new acoustic training and language model training data reduces the word error rate by 2% absolute.

| | baseline | +POS | +WordP | +Expand | +Tune |
|---|---|---|---|---|---|
| manual | 0.6960 | 0.7079 | 0.7109 | 0.7109 | 0.7067 |
| HTK-2 | 0.6746 | 0.6789 | 0.6802 | 0.6804 | 0.6832 |

Table 5: Improvements in Average Precision

We also added several refinements to the IR engine described above. The first was weighting the query terms according to their part of speech, (POS) so proper nouns had the most weight, then common nouns, then adjectives and adverbs, then verbs. Secondly, occurrences of noun-noun or adjective-noun in the stopped query were defined as word-pairs and an additional score was added to those documents which also contained this word pair in the correct order. This meant keeping term position information in the index file. Thirdly we added a small amount of statistical pre-search expansion for the query terms and finally we tuned the constants $b$ and $K$ for our HTK-2 system. The resulting average precision figures, showing some performance advantages for the CU60 queries on the TREC-6 data are given in Table 5.

## 7. CONCLUSIONS

A complete system for spoken document retrieval has been described and the effect of transcription accuracy on retrieval performance has been evaluated. A set of new term error rates has been defined which we feel is a more relevant measure of transcription accuracy for SDR purposes. Future work will include using alternative recognition hypotheses encoded in word lattices for IR and increasing the speed of the transcription engine while retaining high accuracy to allow still larger data sets to be transcribed.

## Acknowledgements

Thanks to Steve Renals for access to Sheffield University's TREC-6 SDR transcriptions. This work is in part supported by an EPSRC grant on "Multimedia Document Retrieval" reference GR/L49611.

## 8. REFERENCES

[1] D Abberley, S Renals, G Cook & T Robinson *The THISL Spoken Document Retrieval System* NIST Special Publication 500-240 Proc. TREC-6 pp. 747, 1997.

[2] J S Garofolo, E M Voorhees, V M Stanford & K Spärck Jones *TREC-6 1997 Spoken Document Retrieval Track Overview and Results* NIST Special Publication 500-240 Proc. TREC-6 pp. 83, 1997.

[3] M F Porter *An algorithm for suffix stripping* Program **14** pp. 130-137, 1980.

[4] S E Robertson & K Spärck Jones *Simple, Proven Approaches to Text Retrieval* Technical Report TR356, Cambridge University Computer Laboratory, May. 1997

[5] P C Woodland, T Hain, S E Johnson, T R Niesler, A Tuerk, E.W.D. Whittaker & S J Young *The 1997 HTK Broadcast News Transcription System* Proc DARPA Broadcast News Transcription and Understanding Workshop, pp. 41-48, Feb. 1998.