The THISL Spoken Document Retrieval System

Steve Renals and Dave Abberley
University of Sheffield
Department of Computer Science
Sheffield S1 4DP, UK
{s.renals,d.abberley}@dcs.shef.ac.uk

ABSTRACT

THISL is an ESPRIT Long Term Research Project focused the development and construction of a system to items from an archive of television and radio news broadcasts. In this paper we outline our spoken document retrieval system based on the ABBOT speech recognizer and a text retrieval system based on Okapi term-weighting . The system has been evaluated as part of the TREC-6 and TREC-7 spoken document retrieval evaluations and we report on the results of the TREC-7 evaluation based on a document collection of 100 hours of North American broadcast news.

Keywords: Multimedia Information Retrieval; Spoken Document Retrieval; Speech Recognition; Broadcast Data.

1 INTRODUCTION

THISL is an ESPRIT Long Term Research project in the area of speech retrieval. It is concerned with the construction of a system which performs good recognition of broadcast speech from television and radio news programmes, from which it can produce multimedia indexing data. The project is concentrating on British and American English applications, with work in progress on a French language system. In particular, the main goal of the project is to develop a system suitable for a BBC newsroom application. The resulting system may be regarded as a "news-on-demand" application in which specific portions of a broadcast may be retrieved in response to a spoken request from the user.

There are two principal approaches to the task of spoken document retrieval. The *phone-based* approach processes the audio data with a lightweight speech recognizer to produce either a phone transcription or a some kind of phone lattice. This data may then be directly indexed or used for word spotting. The *word-based* approach applies a complete large vocabulary speech recognition system to the audio track to produce a word-level transcription; at this point the problem may be treated as standard text retrieval (modulo speech recognizer errors).

The phone-based approach is not restricted by a fixed vocabulary. Since the archiving process only involves phone recognition there is less computational overhead for archiving — although, as discussed in section 6, this may only reduce the computation by around a factor of 2 compared with large vocabulary continuous speech recognition. The phone-based approaches are typically based on indexing overlapping sequences of n phones, where n typically takes values 3 or 4. Ng and Zue [13] have shown that for small spoken document collections this technique can produce average precisions close to that of the reference text, if the perfect phone transcription is known. Working with the output of an automatic phone recognizer results in a relative degradation in performance of around 30%. This approach has also been adopted by Schauble and coworkers [18] and by Smeaton et al [21] who have performed experiments on the TREC-6 spoken document retrieval evaluation and an application based on an archive of RTE news bulletins.

An alternative phone-based approach uses a word spotter, which enables pronunciation constraints to be incorporated. The phone recognizer may be used to produce a phone probability matrix or a phone lattice (graph), which can be used as the input to a word spotting algorithm. Although such algorithms can run many times faster than real-time [4], they are limited by a linear dependence on the size of the archive. However the process may be made considerably more efficient by using an index based on phone sequences to preselect the areas of speech over which the word spotter should be applied. This approach was first suggested by Dharanipragada and Roukos [7] and has also been used by Kraaij et al [12], and may be regarded as a rescoring of the top ranked documents returned by a purely phone-based system. This approach can result in many false alarms; these may be

reduced by including possible "confuser words" in the word-spotter.

In the THISL project we have adopted a word-based approach, similar to that employed by several other groups (eg [2, 9]). This approach requires more computation than phone-based approaches, since a full large vocabulary decoding needs to be applied to the entire archive. However, it enables the constraints of the pronunciation dictionary and language model to be applied. and text retrieval is more robust when applied to words than phone n-grams. Aside from computational considerations, the most frequently cited drawback of this approach is the problem of out-of-vocabulary words. We do not believe that this is a significant problem, and is certainly outweighed by the advantages of the word-based approach. Indeed, of the ad-hoc topics used in the past five TREC evaluations (TRECs 3–7), 9 out of 900 query words were out of vocabulary relative to the 65,000 word vocabulary used in the experiments reported in this paper. This 1% out-of-vocabulary rate corresponds with that we typically observe when recognizing broadcast news data.

In this paper we present the THISL system for spoken document retrieval which is based on the ABBOT large vocabulary continuous speech recognition (LVCSR) system [17] and well-understood probabilistic text retrieval techniques. In section 2, we outline the ABBOT LVCSR system, focusing on those features that make it particularly appropriate for spoken document retrieval. Section 3 describe the text retrieval methods that we used, and we discuss two possible enhancements: the use of multiple transcriptions (section 4) and the use of query expansion (section 5). A series of experiments were carried out as part of our participation in the TREC-7 spoken document retrieval track, and these are described in section 6. Finally, section 7 discusses some conclusions and outlines our current and future research directions in this area.

2 SPEECH RECOGNITION USING ABBOT

We have used the ABBOT LVCSR system developed at the Universities of Cambridge and Sheffield [17]. The four principal components of a probabilistic LVCSR system are the signal processing module (which typically transforms the time domain waveform into a sequence of acoustic feature vectors), the acoustic model (which models phones in terms of the acoustic features), the pronunciation dictionary and the language model (which gives a probability of occurance of any sequence of words).

2.1 CONNECTIONIST ACOUSTIC MODEL

ABBOT differs from most other state-of-the-art LVCSR systems in that it has an acoustic model based on connectionist networks [3]. Although this model may still be interpreted as a type of HMM, it differs from traditional HMMs by directly estimating the posterior probability of each phone given the acoustic features, rather than the likelihood of that phone generating the acoustics. Posterior probability estimation may be performed by a connectionist network (or set of networks) trained to classify phones. In ABBOT, a set of recurrent networks [16] is used. Direct estimation of the posterior probability distribution using a connectionist network is attractive since fewer parameters are required for the connectionist model (the posterior distribution is typically less complex than the likelihood) and connectionist architectures make very few assumptions on the form of the distribution.

Currently, the acoustic model used in the THISL system consists of two recurrent networks with 53 context-independent phone classes (plus silence). One network estimates the phone posterior probability distribution for each frame given a sequence of 12th order perceptual linear prediction features [8]. The other network performs the same distribution estimation with features presented in reverse order (since recurrent networks are time-asymmetric) and the two probability estimates are averaged in the log domain. Each network contains 384 state units, resulting in a total of about 350 000 acoustic model parameters. The 54 context-independent phone models may be expanded to a set of context-dependent phone models, the context classes being arrived at via a decision tree algorithm. A context class network is used for each context-independent phone class, which (when combined with the context-independent phone probabilities) results in a context-dependent phone probability [11]. In the experiments reported in this paper, the system was trained on 100 hours of broadcast news data released by the Linguistic Data Consortium[1]. Twenty-four hours of this data is not transcribed (commercials, local news, etc.), and further sixteen hours was discarded as being below a confidence threshold before training after computing the average log likelihood per frame during a Viterbi alignment. This system is a simplified version of that used by the CU-CON group in the 1997 DARPA evaluation of broadcast news speech recognition systems (hub 4) [5].

The British English system, used for the BBC application, is currently trained on about twenty-four hours of acoustic training data collected and tran-

[1]The first 100 hours of the so-called Hub 4 training data — see the LDC website at http://www.ldc.upenn.edu/

scribed by the BBC Research Department.

2.2 Pronunciation Dictionary

The pronunciation dictionary specifies the finite set of words that may be output by the speech recognizer, and gives at least one pronunciation (ie phone sequence) for each. In the current system, for American English, a dictionary of 65 532 words is used, with a total of about 72 000 pronunciations. The figures are similar for the British English system.

2.3 Language Model

The role of the language model is to estimate the probability $P(w_1 w_2 \dots w_n)$ of a string of words $w_1 w_2 \dots w_n$. This may be decomposed as:

$$P(w_1 w_2 \dots w_n) = P(w_1)P(w_2|w_1)\dots P(w_n|w_{n-1}\dots w_1). \quad (1)$$

If it is assumed that the probability of a word is dependent only on the two preceding words, then (1) may be approximated as:

$$P(w_1 w_2 \dots w_n) = P(w_1)P(w_2|w_1)\dots P(w_n|w_{n-1}, w_{n-2}). \quad (2)$$

This is referred to as a trigram language model. Simple maximum likelihood estimation of trigram probabilities of the form $P(w_n|w_{n-1}, w_{n-2})$ will result in zero probabilities for any trigrams that do not occur in the training data. To prevent this a smoothing technique must be employed. Backing-off involves a portion of the probability mass being reserved for unseen trigrams; this probability mass is split using bigram estimates, and the process may be continued recursively [10].

For the experiments reported here a backed-off trigram language model containing 65 532 unigrams, 7.1 million bigrams and 24.0 million trigrams was estimated from a variety of sources including 132 million words of transcribed broadcast news data and 153 million words of newswire data. The vocabulary was selected by including all the words from the transcription of the acoustic training data, made up to 65,532 words using the most frequent words extracted from the broadcast news text corpus (ignoring common misspellings and obvious text processing errors).

2.4 Search

The search problem in speech recognition may be posed as follows: what is the most probable sequence of word models (or phone models) given the observed acoustics, the acoustic model, the language model and the pronunciation dictionary? Potentially the search space is huge: for example, in the system described above anyone of 65 532 words could start each 16ms. To efficiently evaluate this search space, the Abbot system employs a start-synchronous stack-based search, with substantial pruning of improbable hypotheses [14]. In particular, the search algorithm makes direct use of the posterior probability estimates produced by the neural network acoustic model by pruning all those phones which have an estimated local posterior probability below a threshold. On average, this enables about 70% of the phonetic search space to be pruned at any one time, with a minimal increase in search error.

3 Text Retrieval

In our initial work on spoken document retrieval [1], we used the PRISE text retrieval system[2] developed by NIST. More recently we have developed an Okapi-style testbed system "textbook" probabilistic system, using a stop list, the Porter stemming algorithm and the Okapi term weighting function. Specifically we used the term weighting function $CW(t,d)$ for a term t and a document d given in [15]:

$$CW(t,d) = \frac{CFW(t) * TF(t,d) * (K+1)}{K((1-b) + b * NDL(d)) + TF(t,d)}. \quad (3)$$

$TF(t,d)$ is the frequency of term t in document d, $NDL(d)$ is the normalized document length of d:

$$NDL(d) = \frac{DL(d)}{\overline{DL}}, \quad (4)$$

where $DL(d)$ is the length of document d (ie the number of unstopped terms in d). $CFW(t)$ is the collection frequency weight of term t and is defined as:

$$CFW(t) = \log\left(\frac{N}{N(t)}\right) \quad (5)$$

where N is the number of documents in the collection and $N(t)$ is the number of documents containing term t. The parameters b and K in (3) control the effect of document length and term frequency as usual.

A number of experiments were conducted on a locally derived set of development queries to decide on a suitable stop list and to test the behaviour of running with and without stemming. These experiments clearly indicated that stemming substantially

[2] http://www-nlpir.nist.gov/õver/zp2

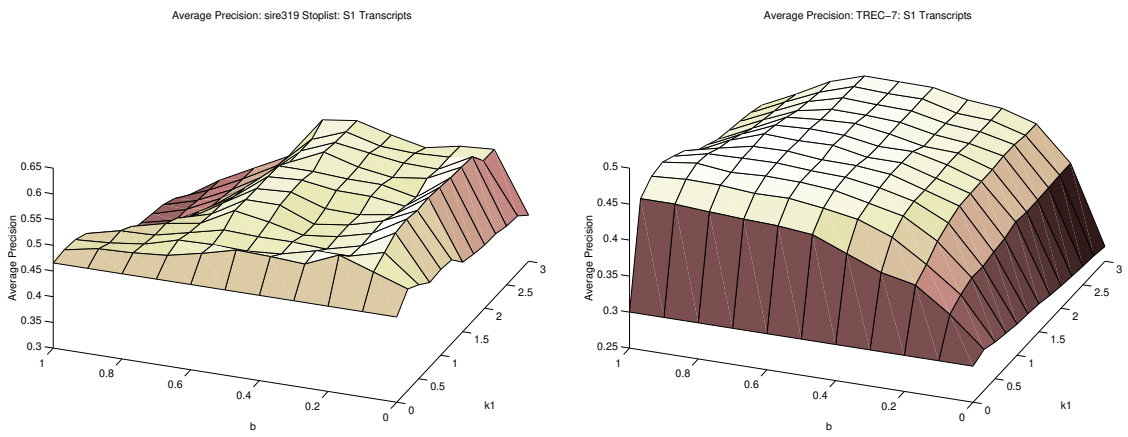

Figure 1: Plot of average precision against term weighting parameters b and K for TREC-7/SDR local development queries (left), and TREC-7/SDR evaluation queries (right).

improved the average precision of the system, and that good performance was achieved using a 379 word stop list (based on the 319 word stop list used by the University of Glasgow at TREC-6 [6]).

Since the task of spoken document retrieval is a little different to text-based ad-hoc retrieval, we investigated the effect of varying the parameters b and K in the term weighting function (3). The results for the development set are shown in figure 1, along with post-evaluation results for the TREC-7 SDR queries. We note that in the development queries there is a ridge of high average precision along $K = 0.25$, which corresponds to a decrease in the significance of TF compared with CFW, which is not present in the evaluation queries. There is also a maximum around $(b,K) = (0.5,1.0)$, for both sets of queries, which (fortunately) were the parameter settings used in our TREC-7 SDR experiments reported in section 6.

The reason for the different behaviour of the two query sets is not clear. Although it may be due to the relatively small task size (around 3000 spoken documents), we also note that our local development queries had many fewer relevant documents per query compared with the evaluation queries (4.5 vs. 17). Support for the latter hypothesis is given by the fact that the parameter landscape for the known-item TREC-6/SDR queries (ie 1 relevant document per query) is most similar to the development set.

4 Multiple Transcriptions

A number of researchers (eg [6, 20]) have taken advantage of the availability of multiple sets of speech recognition transcriptions and merged them to produce improved information retrieval performance. This method was successful because although speech recognizers make errors, different speech recognizers are likely to make different errors. Thus if an important query word has been missed by one recognizer, another one might recognize it correctly so that it does not get omitted from the index.

As mentioned in section 2, the ABBOT acoustic model is based on multiple recurrent networks, which are averaged together at the acoustic frame level. However, it is possible to run separate decodings based on the individual recurrent networks and to merge them together at the transcription level. Experiments were run on the TREC-6 known-item retrieval task using the 379 word stop list but no query expansion. Table 4 shows the results in terms of word error rate (WER), term error rate (TER) and the various TREC-6 IR performance measures. (See section 6 for a definition of TER.)

The table indicates that merging the RNNs at the acoustic probability level (S1) produces better WER/TER and IR performance than either of the individual networks. Despite the inevitably higher TER, merging multiple transcripts seems to produce slightly better IR results than taking their union. The detrimental effects of merging may be partially offset by term frequency weighting. In these experiments, neither merging technique produced clearly better IR performance than the single best set of transcripts (S1), except for the percentage of queries for which the answer was not found.

The results from these experiments are somewhat inconclusive: it is possible that multiple transcripts could be used to enhance retrieval performance but these benefits have yet to be demonstrated unequivocally, and must be offset against the considerable ex-

Transcripts	WER	TER	Mean Rank	Mean Reciprocal	Percentage at Rank 1	Percentage Not Found
R1	–	–	5.85	0.8509	78.7%	0.0%
S1	38.8%	55.4%	11.72	0.7776	74.5%	2.1%
Forward net	43.2%	63.3%	14.33	0.6996	61.7%	2.1%
Backward net	41.7%	61.4%	17.96	0.7091	63.8%	4.3%
Merged fwd+bwd	–	135.9%	14.51	0.7414	68.1%	0.0%
Union fwd+bwd	–	90.3%	18.45	0.7477	68.1%	0.0%
Merged S1+fwd+bwd	–	228.5%	14.40	0.7793	72.3%	0.0%
Union S1+fwd+bwd	–	95.9%	19.77	0.7434	68.1%	0.0%

Table 1: Use of multiple transcriptions derived from ABBOT on the TREC-6 known-item retrieval task. R1 are the reference transcripts, S1 are the transcripts produced by ABBOT using frame-level merging. Forward and backward are the decodings produced by the nets in isolation. The term 'merged' implies the concatenation of two or more sets of transcripts whereas the term 'union' implies the union of sets of transcripts — multiple occurrences of the same term are discarded.

tra resources required to produce the multiple transcriptions (which is why the experiments were not repeated on TREC-7 data).

5 QUERY EXPANSION

If a relevant document does not contain the terms that are in the query, then that document will not be retrieved. The aim of query expansion is to reduce this query/document mismatch by expanding the query using words or phrases with a similar meaning or some other statistical relation to the set of relevant documents. Such a process may have increased importance in spoken document retrieval, since the word mismatch problem is heightened by the presence of errors in the automatic transcription of spoken documents.

An obvious danger in using relevant documents retrieved from a database of automatically transcribed spoken documents is that the query expansion may include recognition errors. This was an experience reported by the INQUERY group in the TREC-6 SDR evaluation [2]. To avoid this problem we retrieved relevant documents from another collection of newswire text. The query expansion algorithm was then applied to the top n documents retrieved from that collection. The resulting expanded query was then applied to the collection of spoken documents.

We used an algorithm based on the local context analysis algorithm of Xu and Croft [22]. The initial query Q is applied to the secondary query expansion collection. The nr top ranked documents are regarded as relevant; the algorithm is not discriminative so no non-relevant documents are required. A query expansion weight, $QEW(Q,e)$ is defined as follows:

$$QEW(Q,e) = \sum_{t \in Q} CFW(t) * \log\left(\frac{\log(AF(e,t)) * CFW(e)}{\log(nr)} + \delta\right) \quad (6)$$

The potential query expansion terms e are simply those terms in the relevant documents. The term $AF(e,t)$ measures the term frequency correlation of two terms e and t across collection of documents d_i:

$$AF(e,t) = \sum_{i=1}^{nr} TF(e,d_i) * TF(t,d_i). \quad (7)$$

The nt possible expansion terms with the largest weights are then added to the original query, weighted as $1/rank$.

In practice the values of nr and nt are maximum limits, since we threshold so that only those documents with a score greater than 0.8 times the score of the top-ranked document are considered, and only those terms with $QEW(Q,e)$ greater than an empirically-determined threshold are added.

In this work we used the June 1997–February 1998 LA Times/Washington Post portion of the TREC/SDR 1998 LM text corpus as the query expansion database. This corpus contains about 13 million words and about 22,000 documents. The parameters nr and nt are clearly dependent on the size of the query expansion collection. Experiments to investigate the dependence on these parameters were carried out on our local development queries, and the results are shown in figure 2. From this we chose parameter values $(nr,nt) = (8,10)$. Figure 3 shows the performance of query expansion using a newswire corpus versus expanding on the target recognizer transcripts.

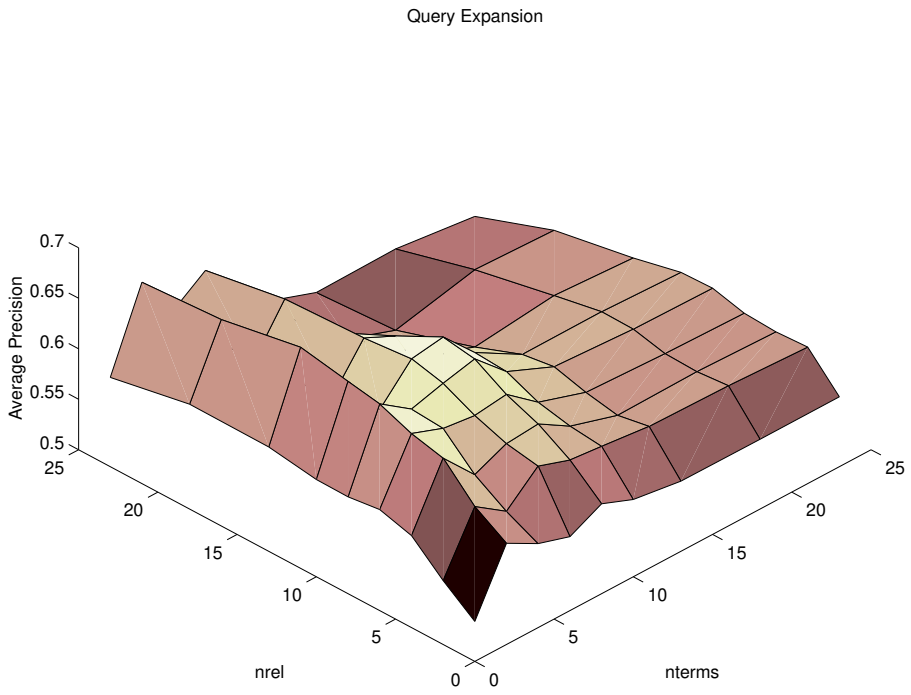

Figure 2: Effect of the query expansion parameters *nr* (maximum number of relevant documents to consider) and *nt* (maximum number of terms to add) on the average precision for our local development queries using ABBOT speech recognizer output. The Jun 1997 – Feb 1998 LA Times/Washington Post portion of the 1998 TREC-7/SDR language model corpus was used as the query expansion collection.

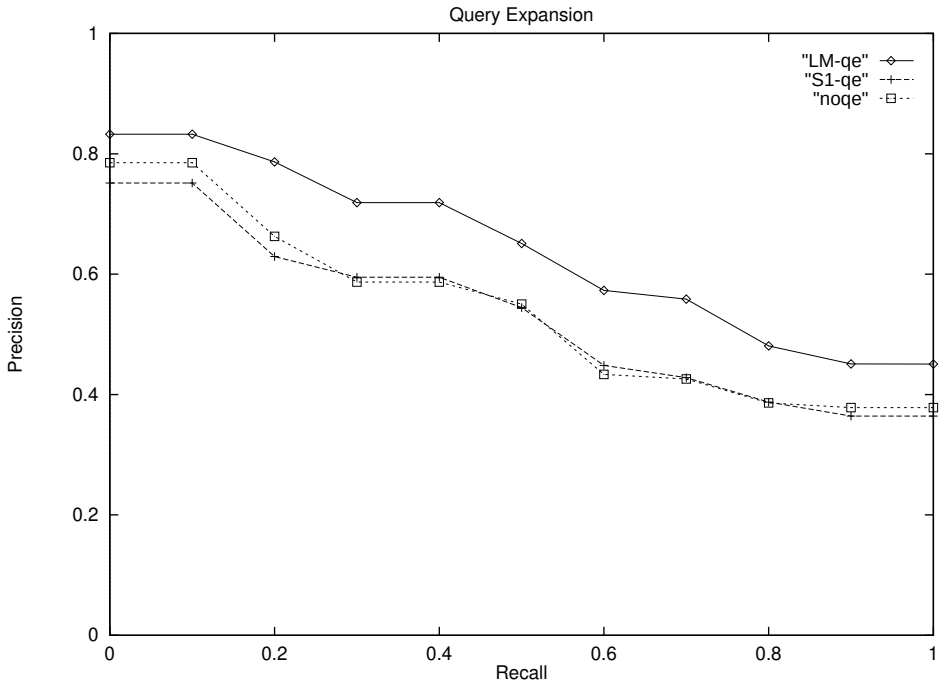

Figure 3: Effect of query expansion on retrieval of recognizer output for local development queries. Query expansion was performed on (1) LA Times/ Washington Post newswire text (LM-qe); (2) the recognizer transcripts that made up the test collection (S1-qe); and (3) no query expansion (noqe).

Condition	WER	TER	Retrieved	Relevant	Rel. Retrieved	AveP	R-P
R1	–	–	17613	390	364	0.4886	0.4583
S1	35.9%	52.2%	18312	390	360	0.4599	0.4485
B1	35.2%	49.5%	18093	390	355	0.4355	0.4562
B2	47.8%	68.3%	18671	390	354	0.3529	0.3347
CR-CUHTK	24.8%	34.0%	18105	390	365	0.4711	0.4469
CR-DERASRU-S1	66.2%	109.3%	17844	390	334	0.3780	0.4164
CR-DERASRU-S2	61.5%	93.7%	17973	390	344	0.4047	0.4016
CR-DRAGON-S1	29.8%	49.2%	18252	390	361	0.4613	0.4372

Table 2: Summary of TREC-7 Spoken Document Retrieval track results for different recognizer conditions, evaluated in terms of word error rate (WER), term error rate (TER) defined in the text, average precision (AveP) and R-precision (R-P). **R1** refers to the reference transcripts; **S1** refers to THISL speech recognition described in the paper; **B1** and **B2** are baseline recognition runs with different levels of pruning using CMU Sphinx-III at NIST; **CR-CUHTK** refers to Cambridge University (HTK) speech recognition; **CR-DERASRU-S1** and **CR-DERASRU-S2** refers to DERA/SRU speech recognition; **CR-DRAGON-S1** refers to Dragon Systems speech recognition.

Note that expanding on the recognizer transcripts is worse than no query expansion.

6 Experiments

6.1 TREC-7 Spoken Document Retrieval

In this section we report the results carried out by the THISL group as part of the TREC-7 Spoken Document Retrieval track. This track involved a collection of 2868 news stories totalling 74 hours of broadcast audio (segmented from a total corpus of 100 hours, and not including commercials, local news, etc.). The audio track was presegmented into stories by hand, and both recognition and retrieval was performed with knowledge of this segmentation. Twenty-three queries were provided by NIST, along with pooled relevance judgments after the evaluation.

6.2 Speech Recognition Results

Using the system described in section 2 we were able to recognize the 74 hours of broadcast news audio date in about seven times real time on an Ultra-1/167MHz (512-1024 Mb RAM), with the computation split approximately equally between the recurrent network-based acoustic model and the LVCSR search algorithm. This implies that there was only a factor of two overhead in performing a word level transcription using 65K vocabulary and trigram language model, compared with phone recognition. However, the memory demands of LVCSR are substantial — our decoder requires a machine with 512Mb RAM — whereas the phone recognizer (essentially the recurrent networks) could run in a couple of megabytes. Running at this speed required a higher degree of pruning, resulting in a relative search error (ie, error resulting from incorrect pruning of the search space) was 10–20%.

The overall average word error rate (WER) of the THISL speech recognition system in this evaluation was 35.9%. We can also use an error metric conditioned on the text retrieval system. The *term error rate* (TER) [9] is given by the following formula:

$$TER = \frac{\sum_{t \in T} \mid R(t) - H(t) \mid}{T} \times 100\% \qquad (8)$$

where $R(t)$ and $H(t)$ represent the number of occurrences of *term* t in the reference and hypothesised transcripts respectively. The set of terms T is calculated after the transcripts have been stopped and stemmed but without taking account of term order. Thus TER gives a more accurate measure than WER of the erroneous terms which will be processed during IR. Additionally, calculating WER is meaningless for merged transcripts (section 4), but TER still provides some information about transcript quality. In conjunction with our submitted system, using a 379 word stop list and Porter stemming the THISL speech recognition system returned a TER of 52.2%.

6.3 Query Processing

Before inputting them to the text retrieval system, the queries were put through several pre-processing operations to normalize their appearance: punctuation was removed, all text was converted to lower case and possible abbreviations/acronyms were expanded to cover alternative transcription possibilities, eg “AIDS” was expanded to “aids a. i. d. s.”. No multiwords or phrases were used in the recognition or retrieval pro-

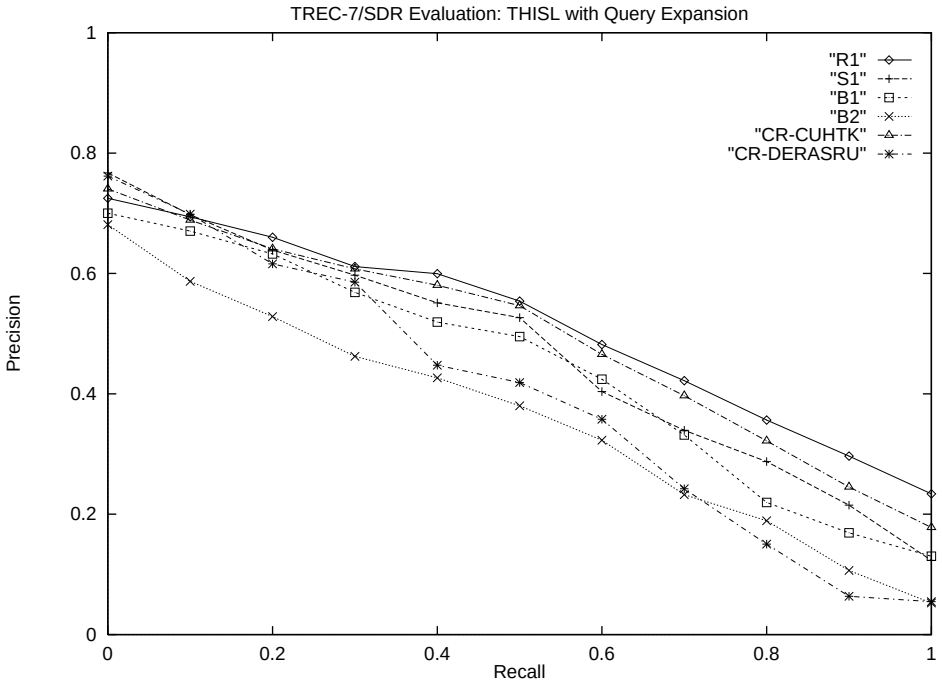

Figure 4: Recall-precision curves of the THISL system running on various transcripts submitted for TREC-7/SDR.

cess. There were three OOV query words: *Montserrat*, *Trie & vs.* (versus). We have previously used a word-spotting system for OOV query words [1], but in these experiments it was not used.

6.4 SDR RESULTS

As well as performing retrieval on the output of our own recognizer, the TREC-7/SDR evaluation permitted retrieval from the transcripts output by the recognizers of other participants. We ran on the recognition output generated by the Cambridge University HTK group, Dragon Systems and DERA/SRU, as well as on the reference transcripts and the output of two baseline recognizers run by NIST. The results were evaluated by word and term error rate and the usual TREC measures of average precision and R-precision, and are shown in table 2.

The recall-precision curves resulting from these runs are shown in figure 4. Figure 5 shows the effect of query expansion on recall and precision for the R1 and S1 conditions. Results for the other speech recognizers are not shown to avoid cluttering the graph, but the effect of query expansion follows a similar trend for those.

Figure 6 shows the relative change due to query expansion for each of the twenty-three queries. As can be seen, query expansion resulted in an improvement or no significant change in average precision for most queries. An example of a query for which the query expansion algorithm proved effective:

60: What information is available on the activities and motivation of intrusive photographers, i.e., the so-called paparazzi?

Original Query: activ avail paparazzi photograph intrus motiv call (AveP = 0.5630)

Expansion Terms: spencer ritz gambino merced editor trespass tabloid (AveP = 0.8589)

A query for which query expansion failed was the following:

62: Find reports of fatal air crashes.

Original Query: air fatal crash (AveP = 0.3520)

Expansion Terms: auto aviat safeti vehicl occup bag jour util (AveP = 0.1893)

7 CONCLUSIONS

We have reported on the development of a spoken document retrieval system for a broadcast news application. Beyond using the straightforward use of a text retrieval system to index the output of a speech

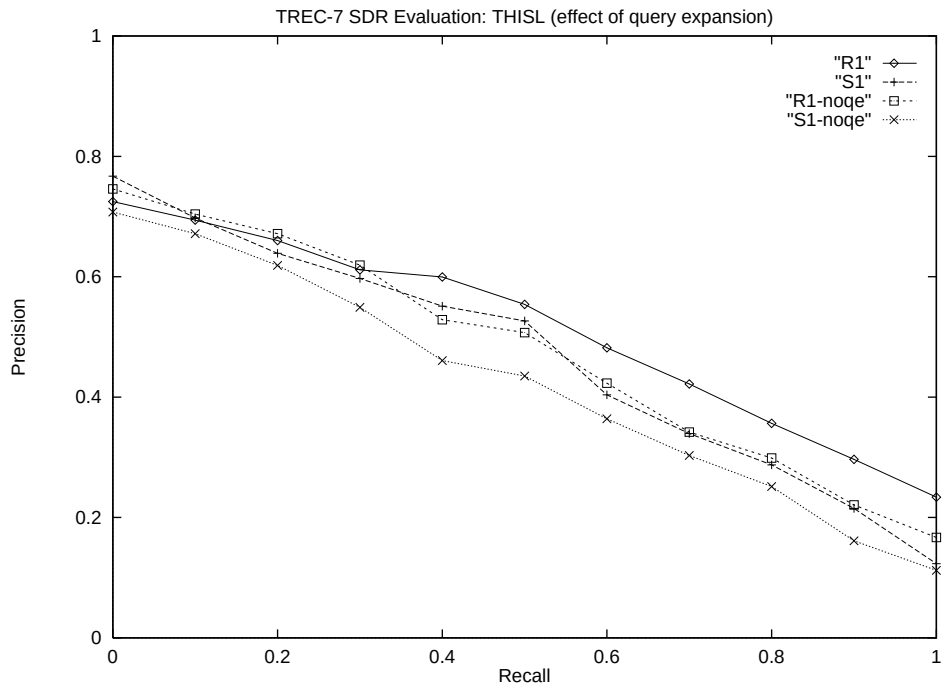

Figure 5: Effect of query expansion on recall-precision for evaluation R1 and S1 conditions (post-evaluation experiment).

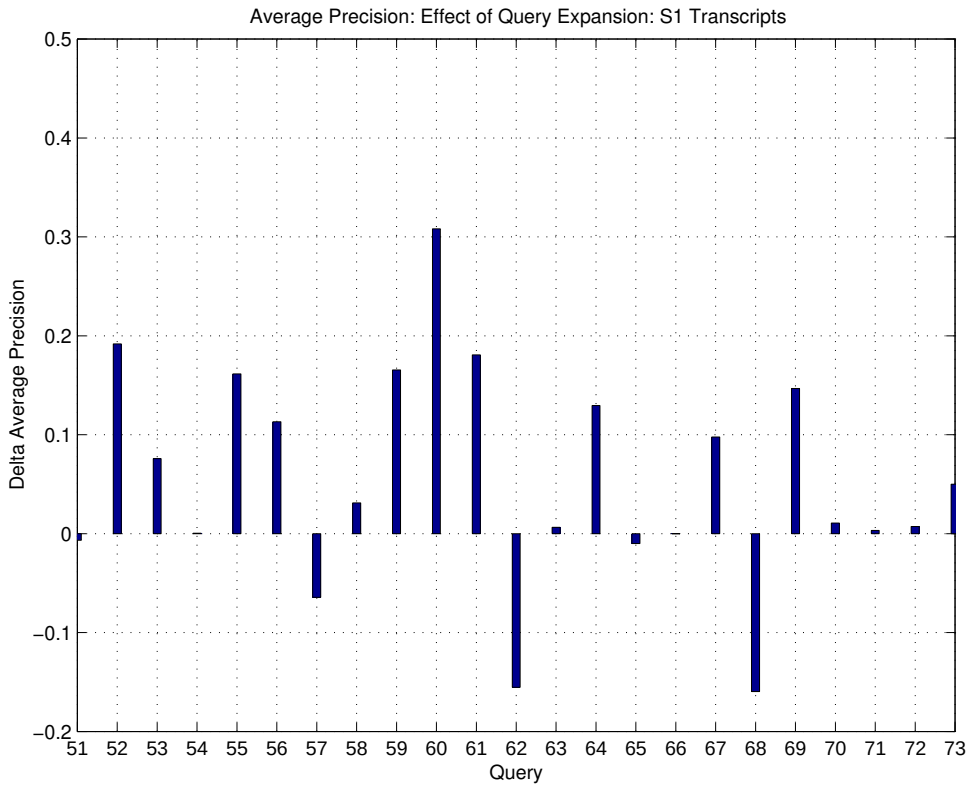

Figure 6: Query-by-query effect of query expansion in terms of change in average precision compared with no query expansion.

recognizer, we have investigated the use of multiple transcription information and query expansion. Document level merging of the possible multiple transcriptions produced by the ABBOT system was not successful in terms of improvements on the TREC-6 known item task. Query expansion, using a secondary collection of newswire data, proved to result in a consistent improvement in average precision of around 10%.

On the 100 hour TREC-7/SDR spoken document collection, our results have indicated that speech recognition systems with word error rates in the region 25–40% are adequate for this task, with only a small degradation from the reference transcripts. There is a correlation with the recognizer word error rate, but there is no clear linear relation between recognition and retrieval performance.

These experiments must be accompanied by the caveat that, in text retrieval terms, we have been working with a very small collection — less than 3000 documents — and experiments to simulate larger collections (eg by corrupting text with a similar number of insertions, deletions and substitutions that a speech recognizer would create) have indicated that difference in average precision between collections of reference transcripts and recognizer output increases with collection size [19]. Although computationally expensive, larger scale experiments in spoken document retrieval are important to test whether this simulated behaviour is accurate. The proposed TREC-8 SDR evaluation, based on 632 hours of broadcast news is a step towards this, as will be the final THISL system based on a large archive of BBC broadcast news.

ACKNOWLEDGMENTS

This work was supported by the ESPRIT Long Term Research Projects THISL (23495) and SPRACH (20077). This work has benefited from collaboration with the partners of the THISL and SPRACH projects, in particular Tony Robinson (Cambridge University and SoftSound) and Gary Cook (Cambridge University).

REFERENCES

[1] D. Abberley, S. Renals, and G. Cook. Retrieval of broadcast news documents with the THISL system. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, pages 3781–3784, Seattle, 1998.

[2] J. Allan, J. Callan, W. B. Croft, L. Ballesteros, D. Byrd, R. Swan, and J. Xu. INQUERY does battle with TREC-6. In *Proc. Sixth Text Retrieval Conference (TREC-6)*, pages 169–206, 1998.

[3] H. Bourlard and N. Morgan. *Connectionist Speech Recognition—A Hybrid Approach*. Kluwer Academic, 1994.

[4] M. G. Brown, J. T. Foote, G. J. F. Jones, K. Spärck-Jones, and S. J. Young. Open-vocabulary speech indexing for voice and video mail retrieval. In *Proc. ACM Multimedia 96*, pages 307–316, Boston, 1996.

[5] G. D. Cook and A. J. Robinson. The 1997 Abbot system for the transcription of broadcast news. In *Proceedings of the 1998 Broadcast News Transcription and Understanding Workshop*, 1998.

[6] F. Crestani, M. Sanderson, M. Theophylactou, and M. Lalmas. Short queries, natural language and spoken document retrieval: Experiments at Glasgow University. In *Proc. Sixth Text Retrieval Conference (TREC-6)*, pages 667–686, 1998.

[7] S. Dharanipragada and S. Roukos. A fast vocabulary independent algorithm for spotting words in speech. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, pages 233–236, 1998.

[8] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *J. Acoust. Soc. Amer.*, 87:1738–1752, 1990.

[9] S. E. Johnson, P. Jourlin, G. L. Moore, K. Spärck-Jones, and P. C. Woodland. The Cambridge University Spoken Document Retrieval System. In *TREC-7 Workshop notebook*, 1998.

[10] S. M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Trans. Acoustics, Speech and Signal Processing*, 35:400–401, 1987.

[11] D. J. Kershaw, M. M. Hochberg, and A. J. Robinson. Context-dependent classes in a hybrid recurrent network-HMM speech recognition system. In *Advances in Neural Information Processing Systems*, volume 8. MIT Press, 1996.

[12] W. Kraaij, J. van Gent, R. Ekkelenkamp, and D. van Leeuwen. Phoneme-based spoken document retrieval. In *Proc. TWLT-14*, 1998.

[13] K. Ng and V. Zue. Phonetic recognition for spoken document retrieval. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, pages 325–328, 1998.

[14] S. Renals and M. Hochberg. Start-synchronous search for large vocabulary continuous speech recognition. *IEEE Trans. Speech and Audio Processing*, in press.

[15] S. E. Robertson and K. Spärck-Jones. Simple proven approaches to text retrieval. Technical Report TR356, Cambridge University Computer Laboratory, 1997.

[16] A. J. Robinson. The application of recurrent nets to phone probability estimation. *IEEE Trans. Neural Networks*, 5:298–305, 1994.

[17] T. Robinson, M. Hochberg, and S. Renals. The use of recurrent networks in continuous speech recognition. In C. H. Lee, K. K. Paliwal, and F. K. Soong, editors, *Automatic Speech and Speaker Recognition – Advanced Topics*, chapter 10, pages 233–258. Kluwer Academic Publishers, 1996.

[18] P. Schauble. *Multimedia Information Retrieval*. Kluwer Academic Publishers, 1997.

[19] M. A. Siegler, M. J. Witbrock, S. T. Slattery, K. Seymore, R. E. Jones, and A. G. Hauptmann. Experiments in spoken document retrieval at CMU. In *Proc. Sixth Text Retrieval Conference (TREC–6)*, pages 291–302, 1998.

[20] A. Singal, J. Choi, D. Hindle, and F. Pereira. AT&T at TREC-6: SDR track. In *Proc. Sixth Text Retrieval Conference (TREC-6)*, pages 227–232, 1998.

[21] A. F. Smeaton, M. Morony, G. Quinn, and R. Scaife. Taiscéalaí: Information retrieval from an archive of spoken radio news. In *Proc. Second European Digital Libraries Conference*, 1998.

[22] J. Xu and W. B. Croft. Query expansion using local and global document analysis. In *Proc. ACM SIGIR*, 1996.