RECOGNITION, INDEXING AND RETRIEVAL OF BRITISH BROADCAST NEWS WITH THE THISL SYSTEM

Tony Robinson (1), Dave Abberley (2), David Kirby (3) and Steve Renals (2)

(1) SoftSound: tony.robinson@softsound.com
(2) Sheffield University: {d.abberley,s.renals}@dcs.shef.ac.uk
(3) British Broadcasting Corporation: david.kirby@rd.bbc.co.uk

ABSTRACT

This paper describes the THISL spoken document retrieval system for British and North American Broadcast News. The system is based on the ABBOT large vocabulary speech recognizer and a probabilistic text retrieval system. We discuss the development of a realtime British English Broadcast News system, and its integration into a spoken document retrieval system. Detailed evaluation is performed using a similar North American Broadcast News system, to take advantage of the TREC SDR evaluation methodology. We report results on this evaluation, with particular reference to the effect of query expansion and of automatic segmentation algorithms.

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 principal objective of the project is to construct a spoken document retrieval system, suitable for a BBC newsroom application. Additionally, we have constructed systems based on North American broadcast news, and a French language version is being developed. In this paper we shall describe the development of both the British and American English systems. Although British English is the main target for our demonstrator, working in American English enables us to evaluate the system performance through the TREC spoken document retrieval track.

The THISL system uses the ABBOT large vocabulary continuous speech recognition (LVCSR) system [1] and well-understood probabilistic text retrieval techniques. Section 2 discusses the overall approach, with the collection of the application specific acoustic and textual data discussed in section 3. Section 4 outlines the speech recognition system detailing changes for the task of British English broadcast news. Section 5 describes the text retrieval methods that we used, with particular attention to the use of query expansion and the development of automatic algorithms to segment streams of broadcast audio into "documents" suitable for text retrieval. The overall implementation of the THISL system is described in section 7 and evaluated in section 8.

2. APPROACH

There are two principal approaches to the task of spoken document retrieval, the *phone-based* approach and the *word-based* approach. In the THISL project we have adopted a word-based approach, similar to that employed by several other groups (eg [2, 3]). 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: text retrieval is more robust when applied to words rather 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 what is typically observed when recognizing broadcast news data.

This work was supported by ESPRIT Long Term Research Project THISL (23495).

3. BRITISH ENGLISH DATA COLLECTION

To cover a reasonably wide range of conditions, speakers and topics, acoustic and textual data for training the British English version was gathered from a variety of BBC News and Current Affairs programmes. In total 45 hours of recorded programmes were transcribed, the majority of which were from television and radio news bulletins but with about 15% from other programmes of a political or financial nature. Transcriptions were carefully checked to ensure they accurately represented the acoustics, as is standard practice. However, we departed from the normal practice of adding fine granularity timing information, say at the end of each sentence or speaker turn, as we found that this was particularly labour intensive. The timing of major changes in acoustic condition were noted but otherwise we only added synchronization marks every five minutes and we further developed our speech alignment software to take the coarse timing information and provide word and phone alignments.

Textual data was acquired from a wider range of sources although still centred on news. Access to the BBC News text database provided material from March 1997 onwards and this was again supplemented with material from related programmes. In total these sources provided about 6.4 million words.

4. SPEECH RECOGNITION USING ABBOT

We have used the ABBOT LVCSR system developed at the Universities of Cambridge and Sheffield [1] and further developed by SoftSound. ABBOT differs from most other state-of-the-art LVCSR systems in that it has an acoustic model based on connectionist networks [4]; here we used two recurrent networks trained on forward-in-time and backward-in-time data (PLP front-end). In this application we use a 64K word pronunciation dictionary, together with a trigram language model.

ABBOT has several characteristics that make it suitable for spoken document retrieval applications including realtime (or close to realtime) performance, decoders with low latency and a simple architecture. Here we outline the development of ABBOT for British English broadcast news; the North American broadcast news system is described in [5].

Acoustic Models: Acoustic models were trained on most of the transcribed corpora. In order to reduce the manual effort in checking transcriptions we filtered the training data using a measure of the confidence that the alignment was in fact the true transcription. The confidence measure chosen was the average log probability of the labelled phone class.

Language Models: For the North American Broadcast News system, language model construction was straightforward, involving the estimation of n-gram language models from text data provided for ARPA/NIST evaluations. There is currently less processed data for the British English system. The trigram language models use

System	WER
baseline system	29.2%
3x real-time	29.0%
with North American LM	30.7%
without cross-sentence	29.4%

Table 1: Overall word error rates by ASR variation

Show	time	date	WER
BBC 1	9pm	8 May 1998	33.0%
BBC 1	6pm	1 Feb 1999	37.7%
BBC 1	1pm	9 Feb 1999	37.1%
Radio 4	6pm	10 Feb 1999	23.4%
Radio 4	6pm	11 Feb 1999	20.6%
Radio 4	6pm	16 Feb 1999	24.1%

Table 2: Word error rates by show

some of the North American text data, together with British English newspaper and newswire data (about 4 million words from Sep–Dec 1998)), transcriptions and scripts from BBC news and current affairs output (about 6 million words from Mar 1997 – Sep 1998) and transcriptions from CNN output (about 8 million words from Sep–Dec 1998).

Search: The LVCSR search space is huge. In the ABBOT system we have adopted stack decoding search strategies embodied in the NOWAY [6] and CHRONOS [7] decoders. We have further developed the CHRONOS decoder for this search task to achieve:

Real-time recognition Using a 450MHz Pentium-II running UNIX we average real-time decoding with a typical memory usage of under 256Mb. This is important for this task as we expect to have over 1000 hours of audio in our final system.

Whole show decoding The efficient memory usage of CHRONOS allows decoding of hour-long shows and so enables the use of online acoustic normalisation as an alternative to the more common per-segment normalisation techniques.

Cross sentence decoding In common with most implementations, our language model contains a special symbol, <s>, to indicate a sentence boundary. Giving this symbol an acoustic realisation of a short period of silence allows the decoder to hypothesise sentence boundaries, and so fit the desired functionality of multiple sentence decoding.

4.1.Speech Recognition Results

Our primary objective is fast, efficient information retrieval. Speech recognition performance is weakly correlated with this goal and in this section we give the word error rate (WER) for various configurations of our system. In many cases we are prepared to accept an increase in WER in order to maximise the overall system performance.

Table 1 shows the WER of the system evaluated on six half-hour BBC news broadcasts. The baseline system was set up to run in real-time, it used the language model described above with online acoustic normalisation instead of segmentation and used cross sentence decoding. The baseline WER is higher than that reported for North American broadcast news system [5], in part because we decode complete broadcasts and score against single hypothesis transcriptions with no flexibility for reasonable variants. The three times real-time system shows that we make only a few more errors in order to run at the speed we desire. We don't expect any of the error rate changes to have a significant effect on information retrieval performance.

More interesting is the show-by-show breakdown of the error rate as given in Table 2. Over the shows we have evaluated, radio news is significantly easier to recognise. Figure 1 plots the error rate throughout a show measured using a 15 second rectangular window. The dashed lines mark the story boundaries and we note that as news topics are introduced by the newsreader we often see

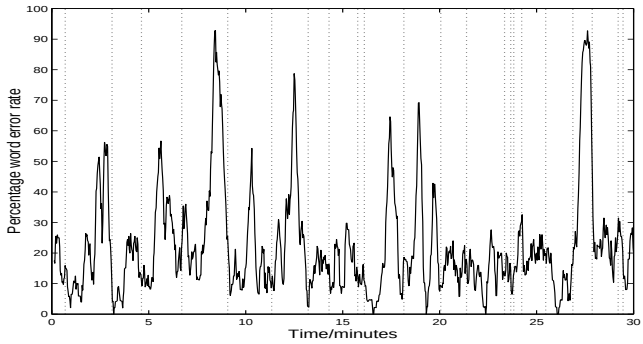

Figure 1: WER over time for Radio 4 News of 10 Feb 1999.

a low WER at the start of a topic. Also we note that there is a very large variation in WER within a topic. This has implications for unsegmented information retrieval and related areas such as audio summarisation where we would like to concentrate on the sections where the speech recognition system is performing well.

5. TEXT RETRIEVAL

The information retrieval component of THISL is based on the bag-of-words probabilistic model. Each document — which is produced by a speech recognizer — is preprocessed using a stop list and the Porter stemming algorithm, and may be represented as a bag of processed terms. We use the Okapi term weighting function [8] to match a term with a document. This is a variant of the usual $tf \cdot idf$ weighting function, with parameters that control the influence of document length and term frequency.

Under the bag of words model, if a relevant document does not contain the terms that are in the query, then that document will not be retrieved. Query expansion may be used 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. This procedure may have even greater importance in spoken document retrieval, since the word mismatch problem is heightened by the presence of errors in the automatic transcription of spoken documents. To avoid the problem of recognition errors corrupting the query expansion process (and to partially compensate for out-of-vocabulary words) we use a secondary corpus of documents from a similar domain that do not contain recognition errors: contemporaneous newswire or newspaper text.

The secondary collection is ranked with respect to the query. Since we have a purely automatic system relevance judgements are not available, so we use a *pseudo-relevance feedback* approach to query expansion, the local context analysis algorithm of Xu and Croft [9]. This algorithm essentially extracts those terms from the top n documents retrieved from the secondary collection which most frequently co-occur with the query terms, using an weighting incorporating the inverse document frequency.

We experimented with this query expansion algorithm on the TREC-7 SDR corpus. Figure 2 shows the effect on the interpolated recall-precision curve for the reference and speech recognition conditions; when evaluated on a query-by-query basis the average precision was increased for 18 of the 23 queries.

6. SEGMENTATION

The TREC SDR evaluations have included hand segmentations of news broadcasts into story units. This is not available for the BBC corpus which was recorded off air, hence we have investigated some simple methods of automatic segmentation.

We have evaluated automatic segmentation algorithms using the TREC-7 SDR corpus since relevance judgements are available. This is a segmented corpus, so the segmentation information was removed by abutting the stories from a show and removing segment

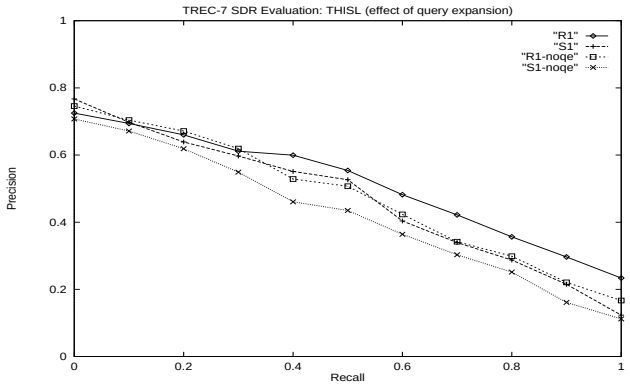

Figure 2: Effect of query expansion on recall-precision for TREC-7 SDR using reference transcriptions (R1) and the output of the ABBOT recognizer (S1).

boundaries. This had the side effect of removing the "gaps" due to unrecognized material such as adverts and sports news. To use the TREC relevance judgements, automatic segments were characterised by a time index (typically the segment mid-point) which could mapped to the appropriate TREC document ID for evaluation.

There has been a substantial amount of work in automatically segmenting documents for text retrieval. Callan [10] and Kaszkiel and Zobel [11] have investigated *passage* retrieval in which documents are broken down into passages typically using document markup or windows of a fixed number of words. Algorithms that automatically segment documents into semantically separate topics have also been investigated recently [12, 13]. Benefits of passage retrieval include the retrieval of the most relevant portions of longer documents, the avoidance of document-length normalization problems and the possibility of more user-friendly interfaces that return the most relevant portion of a document. It has also been claimed that passage retrieval can improve average precision, since it returns short passages with the highest query word density. The principal problems with passage retrieval are the segmentation algorithm, and also the possibility of a dramatic increase in the number of "documents" (ie passages) in the collection.

The situation for spoken data is somewhat different to that for text. Without some kind of prosodic analysis any kind of "document markup" must be at a much coarser level. Also, the average topic length may be much shorter in broadcast news, compared with many text documents.

We have investigated two straightforward approaches to automatic segmentation using windows based on time and number of words. In both cases we have used rectangular windows, of varying lengths and varying degrees of overlap. Initial experiments were carried out using the TREC-7 SDR system, without query expansion. In this case, our standard hand-segmented system resulted in an average precision of 0.4062. Figure 3 shows the average precision for varying window lengths and overlaps, using rectangular windows based on fixed time intervals (left) and fixed word lengths (right). The maximum average precision for both systems is similar, 0.3720 and 0.3757 respectively. This occurs with a relatively short window length (30s and 80 words respectively) and with an overlap of around 50%. The dependence of average precision on window length and overlap appears to be smoother for time-based windowing.

The above experiments were repeated using the local context analysis query expansion algorithm [9], with a secondary corpus consisting of contemporaneous newspaper data (LA Times/Washington Post). Both manual (based on document markup) and automatic (using an 80 word window with 50% overlap) segmentations were employed on this secondary corpus. Since the average document length decreased using the fixed window segmentation, the number of documents used for the pseudo-relevance feedback was increased from 8 to 50.

The results of these experiments (table 3) indicated that the average precision was improved by over 10% for both the manual and automatic segmentations of the spoken documents, when using the document markup segmentation of the secondary corpus. When the secondary corpus was automatically segmented using the fixed length window a further 10% improvement in average precision was observed, rising to 0.50 for the manually segmented spoken documents, and 0.45 for the time-segmented documents. We note that the average precision for time-segmented spoken documents using passage retrieval query expansion is similar to that obtained using manually segemented spoken documents and markup-segmented query expansion.

A side-effect of the automatic segmentation scheme is that adjacent overlapping segments are likely to produce similar scores. Consequently, the list of retrieved documents will contain many segments from the same news item. In an attempt to combine these story fragments, any overlapping segments occurring in the list of retrieved stories are combined into one, larger story. The retrieval score W for the combined story is calculated by applying the following formula:

$$W = \frac{\sum_i^n w_i}{1 + (n-1)\frac{d}{t}}$$

where w_i is the original score for story segment i, n is the number of story segments and t and d are the window length and overlap respectively.

7. BBC NEWS DEMONSTRATION SYSTEM

The current THISL system for BBC news uses the speech recognition and information retrieval strategies discussed above. Query expansion is performed using a secondary collection derived from the British Press Association newswire and we use a time-based rectangular window for automatic segmentation.

The size of the database, and the fact that it would be updated with new programmes each day, required careful consideration of the amount of data that could be processed and handled in practice. It was decided to build the main English language database by taking the six main daily BBC News broadcasts: three each from television and radio channels. This amounts to about 2.5 hours of audio and, although by no means the full output from a newsroom, should cover all the major breaking stories. The database currently consists of over 600 hours of BBC news output, mostly from the period from early 1998 to the present (April 1999), with more complete coverage over the last six months.

8. EVALUATION

The performance of an information retrieval system is very hard to evaluate because the established metrics depend on *relevance assessments* for each of the documents retrieved in response to a query. This information is labour-intensive to collect and is available for the TREC domain but not for the BBC news system. However, there is a standard IR measure which minimises the number of relevance assessments required, the *Precision at Document Number* measure. Hence we have followed a twofold evaluation strategy: evaluation of the equivalent North American broadcast news system within TREC, and precision-oriented evaluation of the BBC system.

Table 3 reports results from the TREC-7 SDR evaluation experiment with and without automatic segmentation. Note that the parameters for the segmentation window were developed on the evaluation set. We note that automatic segmentation only results in a 10% relative reduction in average precision compared with the hand-segmentation. Also, using passage retrieval (with an 80 word window, with 50% overlap) on the secondary query expansion collection results in a small improvement in average precision, compared with using the document boundaries given in that collection. This is consistent with the results reported in [9].

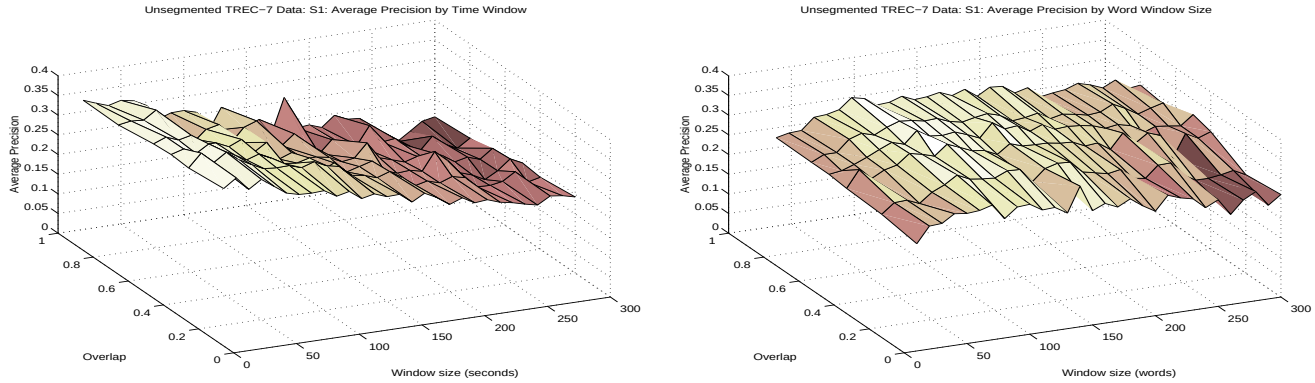

Figure 3: Effect on average precision of automatic segmentation window length.

Query Expansion	Segmentation	Average Precision
No	Manual	0.4062
No	Time	0.3720
No	Words	0.3757
Document	Manual	0.4598
Document	Time	0.4226
Document	Words	0.4254
Passage	Manual	0.5024
Passage	Time	0.4577
Passage	Words	0.4170

Table 3: Average precision for the TREC-7 SDR evaluation data. Conditions included no query expansion and query expansion using a contemporaneous newswire corpus with either passage or document segmentation; and manual segmentation (provided by NIST/LDC) and automatic segmentation based on fixed rectangular time (30s, 60% overlap) or word (80 words, 50% overlap) windows.

The Precision at Document Number is defined as the precision obtained after a given number of documents have been retrieved:

$$\text{Precision at Document No.} = \frac{\text{No. relevant documents retrieved}}{\text{No. documents examined}}$$

This reflects system performance as a user might experience it and it also provides an opportunity to perform some quantitative evaluation of the system. If the number of documents examined is restricted to 10, say, then relevance assessment becomes a manageable task, requiring just 10 judgments per query.

An experiment was conducted by interrogating the BBC News system with the 23 queries from TREC-7[14]. 14 of these had to be modified in some way to make them compatible with the BBC News database to reflect differing news values and the differing time periods covered. The Precision at Document Number results show that approximately half of the first five and four out of the first ten documents returned were relevant (see Table 4). The corresponding figures for the TREC-7 S1 (speech recognition transcripts) and R1 (manual transcripts) runs are included for contrast. The manual transcripts gave slightly better figures than the other two runs but the two experiments are not directly comparable due to differing databases, queries and numbers of relevant documents.

	Precision		
Document Level	BBC News	TREC-7 S1	TREC-7 R1
At 5 Docs	0.5048	0.5130	0.5304
At 10 Docs	0.4143	0.4043	0.4478

Table 4: BBC News performance at document level averages

9. CONCLUSION

We have built a complete system for the recognition, indexing and retrieval of British Broadcast News. What we have learned from the TREC SDR evaluations appears to carry over into the British Broadcast News domain. In addition we have further developed our speech recognition to provide fast processing of whole, unsegmented shows and information retrieval to perform query expansion and to combine overlapping windows to retrieve appropriate audio segments.

REFERENCES

[1] T. Robinson, M. Hochberg, and S. Renals, "The use of recurrent networks in continuous speech recognition," in *Automatic Speech and Speaker Recognition – Advanced Topics* (C. H. Lee, K. K. Paliwal, and F. K. Soong, eds.), ch. 10, pp. 233–258, Kluwer Academic Publishers, 1996.

[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)*, pp. 169–206, 1998.

[3] S. E. Johnson, P. Jourlin, G. L. Moore, K. Sparck Jones, and P. C. Woodland, "The Cambridge University Spoken Document Retrieval System," in *Proc. ICASSP*, 1999.

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

[5] G. Cook, K. Al-Ghoneim, D. Ellis, E. Fosler-Lussier, Y. Gotoh, B. Kingsbury, N. Morgan, S. Renals, T. Robinson, and G. Williams, "The SPRACH system for the transcription of broadcast news," in *Proc. DARPA Broadcast News Workshop*, 1999.

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

[7] T. Robinson and J. Christie, "Time-first search for large vocabulary speech recognition," in *Proc. ICASSP*, 1998.

[8] S. E. Robertson and K. Sparck Jones, "Simple proven approaches to text retrieval," Tech. Rep. TR356, Cambridge University Computer Laboratory, 1997.

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

[10] J. P. Callan, "Passage-level evidence in document retrieval," in *Proc. ACM SIGIR*, pp. 302–309, 1994.

[11] M. Kaszkiel and J. Zobel, "Passage retrieval revisted," in *Proc. ACM SIGIR*, 1997.

[12] M. A. Hearst, "TextTiling: Segmenting text into multi-paragraph sub-topic passages," *Computational Linguistics*, vol. 23, pp. 33–64, 1997.

[13] J. Yamron, I. Carp, L. Gillick, S. Lowe, and P. van Mulbregt, "A hidden Markov model approach to text segmentation and event tracking," in *Proc. ICASSP*, 1998.

[14] S. Renals, D. Abberley, G. Cook, and T. Robinson, "THISL spoken document retrieval at TREC-7," in *Proc. Seventh Text Retrieval Conference (TREC-7)*, 1999.