The THISL SDR System at TREC-8

THE THISL SDR SYSTEM AT TREC-8

Dave Abberley (1), Steve Renals (1), Dan Ellis (2) and Tony Robinson (3,4)

(1) Department of Computer Science, University of Sheffield, UK
(2) ICSI, USA
(3) Department of Engineering, University of Cambridge, UK
(4) SoftSound, UK
Email: {d.abberley, s.renals}@dcs.shef.ac.uk, dpwe@ICSI.Berkeley.EDU, ajr@softsound.com

ABSTRACT

This paper describes the participation of the THISL group at the TREC-8 Spoken Document Retrieval (SDR) track. The THISL SDR system consists of the realtime version of the ABBOT large vocabulary speech recognition system and the THISLIR text retrieval system. The TREC-8 evaluation assessed SDR performance on a corpus of 500 hours of broadcast news material collected over a five month period. The main test condition involved retrieval of stories defined by manual segmentation of the corpus in which non-news material, such as commercials, were excluded. An optional test condition required required retrieval of the same stories from the unsegmented audio stream. The THISL SDR system participated at both test conditions. The results show that a system such as THISL can produce respectable information retrieval performance on a realistically-sized corpus of unsegmented audio material.

1. INTRODUCTION

The TREC-8 test collection was obtained from the TDT-2 corpus and consisted of 902 shows (502 hours) of US broadcast news material covering the period from February to June 1998. The collection contained 21754 individual news items (389 hours of material) with the task being to retrieve the set of stories relevant to each of 50 queries. Two retrieval conditions were specified:

Story Boundary Known (SBK) The SBK runs used a corpus which had been segmented manually into individual news stories, with non-news material being excluded. The definition of non-news material for this purpose included fillers as well as commercials.

Story Boundary Unknown (SBU) The SBU runs reflected the more realistic situation where story boundary information is not known *a priori*. Each news broadcast was to be treated as a continuous audio stream and it was the task of the retrieval system to find the location of the news stories contained within it.

The TREC-8 SDR track was designed to test how SDR systems perform with a much larger document collection than they have been evaluated on previously — the TREC-7 SDR track used only 87 hours of broadcast audio data (2866 stories) [1]. A particular concern was that speech recognition errors would become more dominant as corpus size increased: this problem would be aggravated by a rising out of vocabulary rate caused by the language model becoming progressively out of date over the duration of the corpus. Another concern was to observe the effect automatic segmentation of the corpus would have on retrieval performance.

The THISL[1] spoken document retrieval system consists of the 'real time' version of the ABBOT large vocabulary continuous speech recognizer [2] and the THISLIR text retrieval system [3]. ABBOT is used to transcribe broadcast audio material into text which can be indexed and retrieved by THISLIR. The ABBOT transcriptions can be produced in the order of real time on standard hardware. THISLIR can index and retrieve both segmented and unsegmented news broadcasts.

2. ABBOT SPEECH RECOGNITION

ABBOT is a hybrid connectionist/HMM system [4] which estimates the posterior probability of each phone given the acoustic data at each frame. This differs from traditional recognizers which estimate the *likelihood* that a phone model generated the data. Posterior probability estimation is performed by a set of recurrent networks [5] trained to classify

This work was supported by ESPRIT Long Term Research Projects THISL (23495) and SPRACH (20077).

[1] THISL is an ESPRIT Long Term Research project with the objective of developing a spoken document retrieval system which integrates speech recognition, natural language processing and text retrieval technologies. The main goal of the project is to develop a UK English system suitable for a BBC newsroom application. The TREC SDR evaluation provides an ideal framework to evaluate the performance of the system on a closely related task.

phones. 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. Additionally, this approach enables the use of posterior probability based pruning [6] and is able to provide useful acoustic confidence measures [7]. Decoding is performed by the CHRONOS decoder [8].

THISL produced two sets of speech recognition transcripts for the TREC-8 corpus:

S1 The S1 transcripts were produced by the ABBOT 'real time' system which was used by the SPRACH consortium in the 1998 DARPA/NIST Hub-4 Broadcast News evaluation [2].

S2 The S2 transcripts were produced with an improved acoustic model obtained by merging the acoustic probabilities from the 'real time' system with those produced by an acoustic model using modulation-filtered spectrogram features [10]. These transcripts were not produced in time to be used as an official entry in the evaluation but the results obtained with them have been included here as a contrast condition.

2.1. ACOUSTIC MODELLING

2.1.1. S1 REAL TIME SYSTEM

The acoustic model used by the ABBOT real time system consists of two recurrent networks (RNNs) which estimate *a posteriori* context-independent (CI) phone class probabilities. The phone set contains 54 classes, including silence. One network is used to estimate the phone posterior probability distribution for each frame given a sequence of 12^{th} order perceptual linear prediction features [9]. The other network performs the same distribution estimation but with features presented in reverse order, since recurrent networks are time-asymmetric. The probability streams produced by the two RNNs are averaged in the log domain to produce a final set of probability estimates. The models were trained using the 104 hours of broadcast news training data released in 1997.

2.1.2. S2 SYSTEM

The acoustic model for the S2 system was obtained by log domain merging of the probability estimates produced by the RNNs used in the S1 system with those produced by an acoustic model using modulation-filtered spectrogram features [2].

Modulation-filtered spectrogram (MSG) features were developed to be a representation of speech recognition that is robust to the signal variations caused by reverberation and noise [10, 11]. The robustness is obtained by using a signal-processing strategy derived from human speech perception.

The MSG acoustic model used an MLP containing 8000 hidden units trained on all 200 hours of broadcast news training data downsampled to 4 kHz bandwidth.

2.1.3. LANGUAGE MODELLING

The same backed-off trigram language model [2] was used by both the S1 and S2 systems. Approximately 450 million words of text data was used to generate the model, using the following sources:

- Broadcast News acoustic training transcripts (1.6M words),
- 1996 Broadcast News language model text data (150M),
- 1998 North American News text data:
  LA Times/Washington Post (12M), Associated Press World Service (100M), NY Times (190M).

The models were trained using version 2 of the CMU-Cambridge Statistical Language Model Toolkit [12] using Witten-Bell discounting.

The recognition lexicon contained 65432 words, including every word that appeared in the broadcast news training data. The dictionary was constructed using phone decision tree smoothed acoustic alignments [2].

A *fixed* language model and lexicon constructed from material pre-dating the acoustic data were used throughout the evaluation.

3. TEXT RETRIEVAL

3.1. THISLIR

The THISLIR information retrieval system used for TREC-8 is essentially a "textbook TREC system", using a stop list, the Porter stemming algorithm and the Okapi term weighting function. Specifically, the term weighting function $CW(t,d)$ for a term t and a document d given in [13] was used:

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

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

$$NDL(d) = \frac{DL(d)}{\overline{DL}}, \quad (2)$$

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:

$$CFW(t) = \log\left(\frac{N}{N(t)}\right) \quad (3)$$

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 (1) control the effect of document length and term frequency as usual.

3.2. QUERY EXPANSION

If a relevant document does not contain any of the query terms, then the overall query/document weight (computed using (1)) will be 0, and the document will not be retrieved. This can be a particular problem in spoken document retrieval, owing to the existence of recognition errors, and out-of-vocabulary (OOV) query words. *Query expansion* addresses this problem by adding to the query extra terms with a similar meaning or some other statistical relation to the set of relevant documents.

If words are added to a query using relevant documents retrieved from a database of automatically transcribed audio, then there is the danger that the query expansion may include recognition errors [14]. One way to avoid this problem is through the use of a secondary corpus of documents from a similar domain that does not contain recognition errors. For a broadcast news application, a suitable choice for such a corpus is contemporaneous newswire or newspaper text. A query expansion algorithm may then operate on the relevant documents retrieved from the secondary corpus.

Rather than using a blind relevance feedback approach to query expansion that maintains the term independence assumption which underlies the probabilistic model used for retrieval, we have adopted a method based on the consideration of term co-occurrence. Specifically, we have employed a simplified version of the *local context analysis* (LCA) algorithm introduced by [15]. The query expansion weight $QEW(Q,e)$ for a potential expansion term e and a query Q, across a set of R (pseudo) relevant documents is defined as:

$$QEW(Q,e) = CFW(e) \sum_{t \in Q} CFW(t) \sum_{i=1}^{R} TF(e,d_i) \cdot TF(t,d_i). \quad (4)$$

This approach does not consider distractor (non-relevant, but retrieved) documents. A discriminative term may be included by computing a similar QE weight over a set of distractor documents, combining with (4) using a method such as the Rocchio formula (reviewed by [16]). Experiments have indicated that adding such a discriminative term has a negligible effect. The QE weight (4) is used for ranking potential expansion terms only. Additional weighting can take the form of scaling (1) by $1/rank$.

The query expansion corpus contained about 25 million words and 36000 news stories from the following text sources:

- TREC-7 broadcast news reference transcripts from June 1997 to January 1998 (0.75M words)
- LA Times/Washington Post texts from September 1997 to April 1998 (14.9M words)
- NY Times texts from January 1998 to June 1998 (odd days only) (9.4M words)

After some development work on TREC-7 data, all experiments added a maximum of 15 expansion terms. The manual segmentation of the QE corpus into stories was retained, rather than employing an automatic segmentation into fixed length passages. Development work indicated that there was no significant difference between the schemes, in terms of average precision, but the manual segmentation resulted in an order of magnitude fewer documents to index.

3.3. AUTOMATIC SEGMENTATION

One of the problems which arises when building a practical news on demand application is that radio and TV news recordings do not contain explicit information about when individual news items begin and end, and so some sort of automatic segmentation scheme is required. Segmentation can be attempted at different stages of processing:

1. *Prior information* from programme scripts, etc. can be used if such material is available from the broadcaster. This information is likely to be incomplete and can't allow for the dynamic nature of a live news broadcast, such as when a new story breaks during the programme.

2. *Acoustic information* can also be used to segment a news broadcast. It is possible to detect periods of non-news such as silence, music [17], adverts [18], etc and exclude these from the material to be decoded. Segmentation at this stage has the additional advantage of reducing the amount of material to be decoded (which can be extremely time-consuming when it is non-speech) and reducing the amount of spurious material to be indexed.

3. *Speech recognition transcriptions* are lists of recognized words together with their start and end times. This information can also be used in the segmentation process.

In TREC-8, the only available prior information was the close caption text of the reference transcripts, and the rules of the evaluation forbade its use. No acoustic segmentation was tried due to lack of time for development work. Whilst this decision led to a decoding overhead, experiments on the TREC-7 evaluation suggested that retrieval performance was unlikely to be hit drastically [19]. Consequently, the THISL system in TREC-8 used an automatic segmentation scheme which relied solely on the information provided by the speech recognition transcriptions.

Following the work of Smeaton *et al* [20, 21], a series of experiments were conducted on the TREC-7 dataset using rectangular windows of various lengths and varying degrees of overlap [19]. Window lengths measured in both time and number of words were tried. Relatively short windows worked best but there was not much performance difference between word and time windows. The latter, however, enjoy the advantage of ensuring each document has a maximum time duration and so THISL used 30 second windows with a 15 second overlap for the SBU runs. The document length normalization parameter b was set to zero.

3.3.1. DOCUMENT RECOMBINATION

Each broadcast was segmented into a set of overlapping documents of 30 seconds duration which were then indexed by THISLIR. These documents obviously bore little relation to the actual news stories which would have been obtained by hand segmentation of the corpus, and this created an additional problem. For scoring purposes, a document was identified in terms of a characteristic time within a broadcast, and it was considered to be *relevant* if that characteristic time fell within the time period (defined by manual segmentation) covered by a *relevant* news story. One of the problems with the THISL segmentation scheme is that adjacent overlapping segments are likely to produce similar scores, causing the list of retrieved documents to contain several segments from the same news item. Any such *additional* documents would be scored as *irrelevant*, and so a document recombination and rescoring scheme was devised.

A simple scheme was used. For each query, the top-scoring 4000 documents were retrieved initially. Any documents from the same broadcast which overlapped each other were recombined into one larger document *provided* that their retrieval rank differed by no more than 200 positions. This *200 rule* was introduced to try and prevent low scoring documents from the same broadcast containing 'random hits' of, say, one relatively unimportant query word from being included in the recombined document. The value of 200 was arrived at by conducting a series of tuning runs on the TREC-7 evaluation set which was redecoded (including the non-news portions) for development work. The optimal recombination threshold is likely to be highly corpus and task dependent and merits further empirical examination: a scheme making use of term weighting would also be worth investigating.

The problem of how to rescore the combined documents was also investigated experimentally. Several schemes were tried including using the maximum score from the set of documents to be combined:

$$W_{\max} = \max_i^n w_i \tag{5}$$

reestimating the Okapi score for the combined document (updating CFW, but not accounting for the overlap between adjacent documents), and other methods with less obvious theoretical justification. On the TREC-7 development data, the best performing rescoring formula proved to be:

$$W_{\text{DERB}} = \frac{\sum_i^n w_i}{1 + (n-1)\frac{d}{t}} \tag{6}$$

where W is the retrieval score for the combined document, w_i is the original score for document i, n is the number of document segments to be combined and t and d are the window length and overlap respectively. This formula — known locally as the *DERB factor* — was arrived at somewhat accidentally[2] for our UK English system [22] and has the effect of boosting the score of a combined document relative to that of a standalone document. It does not require term frequency information to obtain the new score, and hence can be implemented by post-processing the raw retrieval output.

Subsequent experiments on TREC-8 evaluation data showed that using the maximum score from the set of documents to be combined produced an improvement in average precision (see Section 4.2.3).

3.3.2. BROADCAST SEGMENTATION

The THISL automatic segmentation can be summarized as follows:

1. Entire news broadcast decoded into a *stream* of text.
2. Text stream broken into *documents* using a fixed length rectangular window of 30 seconds with a 15 second overlap.
3. Resulting documents are indexed by THISLIR.
4. At retrieval time, the 4000 top-scoring stories are retrieved. Overlapping documents from the same show are combined into *stories* if retrieval rank difference < 200.
5. Retrieval score adjusted for each story using Equation 6.

3.4. PARAMETER SETTINGS

A locally developed 379 word stop list and the Porter stemming algorithm were used.

The term weighting parameter settings for the SBU and SBK runs are given in Table 1. The SBK parameters have been changed slightly from their TREC-7 settings owing to the larger query expansion database and as a result of experience with our UK English system [22]. Note that the

[2] Thanks to Sue Johnson for pointing this out!

Parameter	SBK	SBU
b	0.7	0.0
K	1.5	1.5
QE-b	0.5	0.5
QE-K	0.25	0.25
QE-nt	15	15
QE-nr	10	10

Table 1: Parameter settings for TREC-8 SDR SBK and SBU runs.

document length parameter *b* was set to zero for the SBU run because the automatic segmentation scheme inherently performs approximate document length normalization (see Section 3.3.1).

3.5. QUERY PREPARATION

The text queries were preprocessed before being input to THISLIR by removing punctuation, converting to lower case and expanding numbers and abbreviations/acronyms to make the query more similar to speech recognizer output (eg, *1998* $\rightarrow$ *nineteen ninety eight*, *G-7* $\rightarrow$ *G seven*).

4. EVALUATION RESULTS

4.1. SPEECH RECOGNITION

Nominal word error rates (WERs) were estimated on a 10 hour subset of the corpus. The ABBOT S1 system produced a WER of 32.0%, and the S2 system improved this to 29.2%. The 'real time' S1 decodings were produced in approximately 3 $\times$ real time on a variety of standard hardware. Note that this is the overall *average* figure and the decoding speed of a given broadcast will vary with machine performance. Further, entire news shows were decoded, and the speed of the search phase will have been compromised due to the decoder having to transcribe material such as commercials which it had not been trained on. Table 2 gives a breakdown of the time taken for the different stages of decoding. The S2 system was slightly slower due to the extra processing involved. Dedicated hardware was used at the S2 acoustic modelling stage.

No multiwords or phrases were used in the recognition or retrieval process. OOV words were not a significant problem; as usual, there was one OOV word in the TREC queries (*Filo*), together with a text processing problem (*II's* (as in "Pope John Paul II's") was not expanded to "the second's").

Process	$\times$ real time S1	S2
Feature extraction	0.1	0.2
Acoustic modelling	0.2	0.5
Search	2.8	2.8
Total	3.1	3.5

Table 2: Average time taken at stages of the decoding process.

4.2. INFORMATION RETRIEVAL

4.2.1. STORY BOUNDARY KNOWN (SBK) CONDITION

Table 3 shows the results for the SBK runs with THISLIR for the different sets of transcriptions. Average precision is seen to decrease slightly as Word Error Rate (WER) increases (Figure 1). The S2 run did not produce an improvement in average precision relative so S1 despite the improved WER.

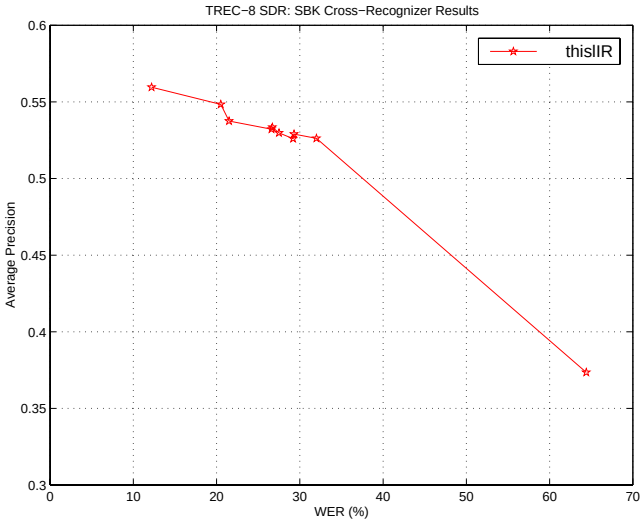

Figure 1: thislIR: SBK Average Precision as a function of WER

4.2.2. STORY BOUNDARY UNKNOWN (SBU) CONDITION

Table 4 shows the results for the SBU runs. Once again, average precision tends to decrease as WER increases. The improved error rate of the S2 run produced a 1% improvement in average precision. Figure 2 illustrates the trend graphically.

Average precision for the SBU runs is about 10% lower (in absolute terms) than for the corresponding SBK runs. Although this is a considerable loss of performance, the

SBK Run	WER	Retrieved	AveP
shef-r1	12.2%	1653	0.5596
shef-cr-cuhtk-s1	20.5%	1638	0.5484
shef-cr-limsi-s1	21.5%	1613	0.5375
shef-cr-cuhtk-s1p1	26.6%	1621	0.5322
shef-b2	26.7%	1587	0.5335
shef-b1	27.5%	1590	0.5298
shef-s2	29.2%	1609	0.5260
shef-cr-att-s1	29.3%	1622	0.5290
shef-s1	32.0%	1594	0.5262
shef-cr-cmu-s1	64.4%	1299	0.3735

Table 3: Summary of results for Story Boundary Known condition. *WER* is word error rate, *Retrieved* is the number of relevant documents retrieved out of a total of 1818, *AveP* is the average precision.

information retrieval capability of the system is still quite respectable with, on average, over 50% of the top 10 documents retrieved being relevant. This is an encouraging result at this stage in the development of automatic segmentation schemes.

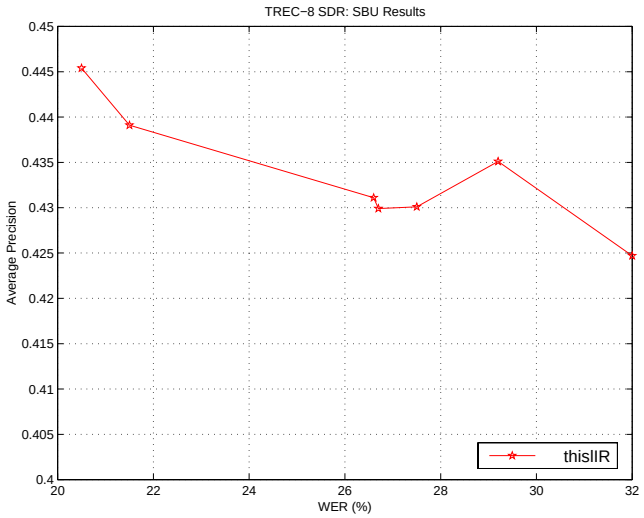

Figure 2: thislIR: SBU Average Precision as a function of WER

4.2.3. *EFFECT OF RESCORING METHOD*

Table 5 compares the DERB (Equation 6) and MAX (Equation 5) rescoring methods (see Section 4.2.3). On the TREC-8 evaluation set, MAX rescoring gives up to a 1% increase in average precision and a small decrease in the number of relevant documents retrieved[3]. It is interesting to note that the increases vary depending on the speech recognition transcripts used (see Figure 3). There is obviously much scope for experimentation to find the optimum rescoring formula.

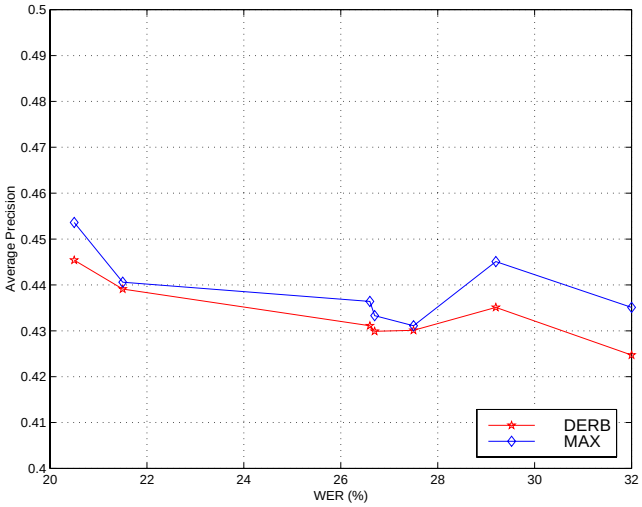

Figure 3: thislIR: SBU Average Precision variation with rescoring method

4.2.4. *EFFECT OF NON-NEWS MATERIAL*

The segmentation procedure described in Section 3.3 made no attempt to exclude non-news material. It is interesting to estimate what effect this had on information retrieval performance. Table 6 compares the performance of the THISLIR SBU system on the shef-s1 and shef-s1u transcripts. The shef-s1 run represents the 'perfect case' where all non-relevant material (and *only* non-relevant material) has been removed. The figures show that indexing up non-relevant material such as commercials, and also fillers, causes a relatively modest 1.3% loss in average precision. This suggests that most of the performance loss associated with auto-

[3]The reverse was true on the TREC-7 data used for tuning experiments.

SBU Run	WER	Retrieved	AveP
shef-cru-cuhtk-s1u	20.5%	1458	0.4454
shef-cru-limsi-s1u	21.5%	1442	0.4391
shef-cru-cuhtk-s1p1u	26.6%	1455	0.4311
shef-b2u	26.7%	1386	0.4299
shef-b1u	27.5%	1393	0.4301
shef-s2u	29.2%	1418	0.4351
shef-s1u	32.0%	1393	0.4247

Table 4: Summary of results for Story Boundary Unknown condition. *WER* is word error rate, *Retrieved* is the number of relevant documents retrieved out of a total of 1818, *AveP* is the average precision.

		DERB		MAX	
SBU Run	WER	Retrieved	AveP	Retrieved	AveP
shef-cru-cuhtk-s1u	20.5%	1458	0.4454	1443	0.4536
shef-cru-limsi-s1u	21.5%	1442	0.4391	1421	0.4406
shef-cru-cuhtk-s1p1u	26.6%	1455	0.4311	1441	0.4364
shef-b2u	26.7%	1386	0.4299	1374	0.4333
shef-b1u	27.5%	1393	0.4301	1375	0.4311
shef-s2u	29.2%	1418	0.4351	1401	0.4451
shef-s1u	32.0%	1393	0.4247	1387	0.4351

Table 5: Comparison of document rescoring methods for the SBU task. DERB refers to the equation presented in Section 4.2.3. MAX refers simply to the rescoring of a recombined document by giving it the highest score of all the individual documents.

matic segmentation is due to the mismatch between the real story boundaries and those defined automatically.

5. CONCLUSIONS

1. The TREC-8 SDR track evaluated current SDR technology on a substantial corpus of broadcast news material. The results show that information retrieval performance was not significantly affected by the size of the corpus or the increased number of out of vocabulary words caused by the language model becoming out of date. In general, problems caused by transcription errors are largely offset by techniques such as query expansion.

2. Automatic segmentation of the corpus with a very simple algorithm resulted in a 10% absolute degradation in average precision. Although this is a considerable loss of performance, the information retrieval capability of the system is still quite respectable with, on average, over 50% of the top 10 documents retrieved being relevant.

6. REFERENCES

[1] J. Garofolo, E. Voorhees, C. Auzanne, and V. Stanford, "Spoken document retrieval: 1998 evaluation and investigation of new metrics," in *Proc. ESCA ETRW Workshop Accessing Information in Spoken Audio*, (Cambridge), pp. 1–7, 1999.

[2] A. J. Robinson, G. D. Cook, D. P. W. Ellis, E. Fosler-Lussier, S. J. Renals, and D. A. G. Williams, "Connectionist Speech Recognition of Broadcast News." Submitted to Speech Communication.

[3] 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.

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

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

[6] S. Renals and M. Hochberg, "Start-synchronous search for large vocabulary continuous speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 7, pp. 542–553, 1999.

SBU Run	Retrieved	AveP
shef-s1u	1393	0.4247
shef-s1 (no adverts)	1415	0.4376

Table 6: Story Boundary Unknown condition: effect of indexing non-news material.

[7] G. Williams and S. Renals, "Confidence measures from local posterior probability estimates," *Computer Speech and Language*, vol. 13, pp. 395–411, 1999.

[8] T. Robinson and J. Christie, "Time-first search for large vocabulary speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, (Seattle), pp. 829–832, 1998.

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

[10] B. Kingsbury, *Perceptually-inspired Signal Processing Strategies for Robust Speech Recognition in Reverberant Environments*. PhD thesis, Dept. of EECS, UC Berkeley, 1998.

[11] B. E. D. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Communication*, vol. 25, pp. 117–132, 1998.

[12] P. Clarkson and R. Rosenfeld, "Statistical language modeling using the CMU-Cambridge toolkit," in *Eurospeech-97*, pp. 2707–2710, 1997.

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

[14] 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.

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

[16] D. K. Harman, "Relevance feedback and other query modification techniques," in *Information Retrieval: Data Structures and Algorithms* (W. B. Frakes and R. Baeza-Yates, eds.), ch. 11, pp. 241–263, Prentice Hall, 1992.

[17] G. Williams and D. Ellis, "Speech/music discrimination based on posterior probability features," in *Eurospeech-99*, 1999.

[18] S. E. Johnson, P. Jourlin, K. S. Jones, and P. C. Woodland, "Spoken document retrieval for trec-8 at cambridge university," in *Proc. TREC-8*, 1999.

[19] D. Abberley, D. Kirby, S. Renals, and T. Robinson, "The THISL broadcast news retrieval system," in *Proc. ESCA ETRW Workshop Accessing Information in Spoken Audio*, (Cambridge), pp. 14–19, 1999.

[20] 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.

[21] G. Quinn and A. Smeaton, "Optimal parameters for segmenting a stream of audio into speech documents," in *Proc. ESCA ETRW Workshop Accessing Information in Spoken Audio*, (Cambridge), pp. 96–101, 1999.

[22] T. Robinson, D. Abberley, D. Kirby, and S. Renals, "Recognition, indexing and retrieval of british broadcast news with the thisl system," in *Eurospeech-99*, 1999.