SPOKEN DOCUMENT RETRIEVAL FOR TREC-9 AT CAMBRIDGE UNIVERSITY

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

†Cambridge University Engineering Department,Trumpington Street, Cambridge, CB2 1PZ, UK.
Email: {sej28,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 presents work done at Cambridge University for the TREC-9 Spoken Document Retrieval (SDR) track. The CU-HTK transcriptions from TREC-8 with Word Error Rate (WER) of 20.5% were used in conjunction with stopping, Porter stemming, Okapi-style weighting and query expansion using a contemporaneous corpus of newswire. A windowing/recombination strategy was applied for the case where story boundaries were unknown (SU) obtaining a final result of 38.8% and 43.0% Average Precision for the TREC-9 short and terse queries respectively. The corresponding results for the story boundaries known runs (SK) were 49.5% and 51.9%. Document expansion was used in the SK runs and shown to also be beneficial for SU under certain circumstances. Non-lexical information was generated, which although not used within the evaluation, should prove useful to enrich the transcriptions in real-world applications. Finally, cross recogniser experiments again showed there is little performance degradation as WER increases and thus SDR now needs new challenges such as integration with video data.

1. INTRODUCTION

With the ever-increasing amount of digital audio data being produced, it is becoming increasingly important to be able to access the information contained within this data efficiently. Spoken Document Retrieval (SDR) addresses this problem by requiring systems to automatically produce pointers to passages in a large audio database which are potentially relevant to text-based queries. The systems are formally evaluated within TREC using relevance assessments produced by humans who have listened to the audio between previously established manually-defined "story" boundaries. A transcription generated manually is also provided for a reference run to give an approximate upper-bound on expected performance.

The natural way to allow easy indexing and hence retrieval of audio information is to represent the audio in a text format which can subsequently be searched. One such method is to represent the speech present in the audio as a sequence of sub-word units such as phones; generate a phone sequence for the text-based query; and then perform fuzzy matching between the two. (see e.g. [4, 17]) The fuzzy phone-level matching allows flexibility in the presence of recognition errors and out of vocabulary (OOV) query words can potentially find matches. However, this approach still requires a method of generating phone sequences from the query words (usually a dictionary); it cannot easily use many standard text-based approaches, such as stopping and stemming; and performance on large scale broadcast news databases, such as those used within the TREC-SDR evaluations is generally poor[8].

With the recent improvements in the performance and speed of large vocabulary continuous speech recognition (LVCSR) systems, it is possible to produce reasonably accurate word based transcriptions of the speech within very large audio databases. This allows standard text-based approaches to be applied in retrieval, and means that a real user could easily browse the transcripts to get an idea of their topic and hence potential relevance without needing to listen to the audio. (see e.g. [27]). The inclusion of a language model in the recogniser greatly increases the quality of the transcriptions over the phone-based approach, and the overall performance of word-based systems has outperformed other approaches in all previous TREC-SDR evaluations [8]. OOV words do not currently seem to present a significant problem provided that suitable compensatory measures are employed [28] and rolling language models have been investigated (see e.g. [3]) as a way to adapt to changing vocabularies as the audio evolves.

Several methods to compensate for the errors in the automatically generated transcriptions have been devised. Most of these use a contemporaneous text-based news-wire corpus to try to add relevant non-erroneous words to the query (e.g. [1, 12]) or documents (e.g. [22, 23, 12]) although other approaches are also possible (e.g. the machine-translation approach in [5]). These methods have proven very successful even for high error rate transcriptions [16], so the focus of SDR has generally switched to trying to cope with continuous audio streams, in which no "document" boundaries are given[1]. This story-boundary-unknown (SU) task is the main focus of the TREC-9 SDR evaluation.

∗ Now Sue Tranter, Dept. of Engineering Science, Oxford, OX1 3PJ, UK : sue.tranter@eng.ox.ac.uk
∗∗ Now at Laboratoire d'Informatique de l'Universite d'Avignon : pierre.jourlin@lia.univ-avignon.fr

[1] Or at least where topic boundaries are not available within the global boundaries of a newscast.

Our overall approach involves generating a word-level transcription and dividing it into overlapping 30 second long windows. Standard stopping, stemming and Okapi-weighting are used during retrieval with query expansion from a contemporaneous newswire collection, before merging temporally close windows to reduce the number of duplicates retrieved.

This paper describes the Cambridge University SDR system used in the TREC-9 SDR evaluation. Sections 1.1 and 1.2 describe the tasks and data for the evaluation in more detail. The problem of extracting non-lexical information from the audio which may be helpful for retrieval and/or browsing is addressed in section 2 and the transcriptions used are described in section 3. Development for the SU runs is given in section 4, with results from the final system on all transcriptions and query sets given in section 5. The effects of using non-lexical information in retrieval are investigated in section 6 and a contrast for the case where story boundary information is known (SK) is given in section 7. Finally conclusions are offered in section 8.

1.1. Description of TREC-9 SDR Tasks

The TREC-9 SDR evaluation [6] consisted of two tasks. For the main story-boundary-unknown (SU) task, the system was given just the audio for each news episode (e.g. entire hour-long newscasts) and had to produce a ranked list of episode:time stamps for each text-based query. The scoring procedure involved mapping these stamps to manually defined story-IDs, with duplicate hits being scored as irrelevant, and then calculating Precision/Recall in the usual way[2].

The two differences from this task to the TREC-8 SDR SU evaluation task [8, 12] are firstly that for TREC-9, all the audio was judged for relevance (including e.g. commercials) and secondly that non-lexical information (such as the bandwidth/ gender /speaker-ID, or the presence of music etc.) that was automatically detected by the speech recognition system could be used in addition to the word-level output at retrieval time. A contrast run (SN) was required without the use of the non-lexical information, if it had been used within the SU run, to allow the effect of this additional information to be seen.

Another contrast run where manually-defined story boundaries were provided (SK) allowed the degradation from losing the story boundary information to be evaluated. This is the same as the primary task in the TREC-8 SDR evaluation. Sites had to run on their own transcriptions (`s1`), a baseline provided by NIST (`b1`) and the manually-generated reference (`r1`)[3].

1.2. Description of Data

The audio data for the document collection was the same as that used in the TREC-8 SDR evaluation, namely 502 hours (˜4.5M words) from 902 episodes of American news broadcast between February and June 1998 inclusive. The SK runs took a subset of ˜3.8M words divided into 21,754 manually defined "stories" to give an average document length of ˜170 words.

The queries used for development (TREC-8) and evaluation (TREC-9) are described in Table 1. Two sets of queries were used, namely *short* (corresponding to a single sentence) and *terse* (approximately 3 key words). The query sets corresponded to the same original information needs and thus the same relevance judgements were used in both cases. The introduction of terse queries was new for TREC-9, and was intended to model the keyword-type query used in many WWW search engines. Since there were no existing terse development queries, *terse forms of the TREC-8 queries were developed in house and thus are not the same as those used by other sites.*

	Dev (TREC-8)	Eval (TREC-9)
Num. Queries	49	50
Ave. # Words in Query	13.6 (s) 2.4 (t)	11.7 (s) 3.3 (t)
Ave. # Distinct Terms per Q.	6.6 (s) 2.3 (t)	5.6 (s) 2.9 (t)
Ave. # Rel Docs	37.1	44.3

Table 1: Properties of query and relevance sets.(s=short t=terse)

The contemporaneous parallel text corpus used for query and document expansion consisted of 54k newswire articles (36M words) from January to June 1998. Although significantly smaller than that used by some other sites (e.g. 183k articles in [24]), in previous work we found that increasing the parallel corpus size to approximately 110k articles did not help performance [16]. The corpus, summarised in Table 2, consisted of the (unique) New York Times (NYT) and 20% of the Associated Press (APW) articles from the TREC-8 SDR Newswire data enhanced with some LA Times/Washington Post (LATWP) stories and was evenly distributed over the whole time period.

Source	LATWP	NYT	APW	Total
Num. Stories	15923	20441	17785	54149
Ave. # Words in Doc.	685	885	385	662

Table 2: Description of the Parallel Corpus.

2. GENERATING NON-LEXICAL INFORMATION

Audio contains much more information than is captured simply by transcribing the words spoken. For example, the way things are said, or who said them can be critical in understanding dialogue, and many non-speech events (such as music, applause, sudden noises, silence etc.) may also help the listener follow what was recorded. Current speech recognisers can automatically recognise many of these things, such as the speaker ID or gender (e.g. [13]) and the presence of music, noise and silence etc. (e.g. [21]), but the speech-recognition-transcription (SRT) format used in the SDR evaluations does not support the inclusion of such additional information. For TREC-9 a new Segmentation Detection Table (SDT) file was allowed [6], which represented various audio phenomena found during recognition in a text-based format which could be used at retrieval time.

[2] Precision and Recall were calculated with respect to whole stories, rather than a more natural passage-based approach for logistic reasons.

[3] See section 3 for more details.

There are two main uses for such non-lexical information, namely to increase retrieval performance and to help navigation/browsing in real SDR applications. The TREC-9 SDR evaluation only allowed the former to be properly evaluated, but the latter is equally important in real world applications, and tags should not be thought to be irrelevant just because they were not used in the retrieval stage of the system [18].

Non-lexical information can be used to help SU retrieval in two main ways. Firstly some information about broadcast structure including potential locations of commercials and story boundaries can be postulated from audio cues such as directly-repeated audio sections, changes in bandwidth/speaker or the mean energy in the signal. Secondly properties such as the presence of music, background noise or narrowband speech can be used to identify portions of transcription which are potentially less reliable than normal.

Table 3 shows the tags generated, whilst the next section explains how these were produced and section 6 discusses their effect on retrieval performance.

Tag Number	(high)-Energy 19,882	Repeat 7,544	Commercial 5,194
Segment 142,914	Gender 57,972	Bandwidth 49,542	Nospeech 15,700

Table 3: Non-lexical tags generated for TREC-9.

2.1. Segment, Gender, Bandwidth and Nospeech

The first stage of our speech recognition system consists of an audio segmenter. Initially the data is classified into wideband speech, narrowband speech or pure music/noise, giving the `bandwidth` and `nospeech` tags respectively. The labelling process uses Gaussian mixture models and incorporates MLLR adaptation. A gender-dependent phone recogniser is then run on the data, and the smoothed `gender` change points and silence points are used in the final segmentation, hence generating the `segment` tags. More details can be found in [12] and [11].

2.2. Energy

Signal energy can help to indicate the presence of commercials. The average normalised log energy (NLE)[4] for the TREC-7 and January TDT-2 data, given in Table 4, shows that in general commercials have a higher mean energy content than news.

	TREC-7 data			January TDT-2 data	
Br.	Story	Filler	Comm.	News	Comm.
ABC	-2.82	-2.82	-1.95	-2.98	-2.22
CNN	-2.22	-2.21	-1.69	-2.27	-2.08
PRI	-2.40	-2.63	-1.84	-2.61	-2.48

Table 4: Average normalised log-energy for TREC-7 and January TDT-2 data for Stories, Fillers and Commercials.

By windowing the audio and comparing the NLE for each 5s window to a threshold, it is possible to generate a crude indicator of where commercials might be occurring. Imposing a minimum length restriction on the postulated commercials can be used to reduce the false alarm rate. Table 5 shows the results of applying such a system on the development (January TDT-2) and test (TREC-9) data. Whilst the method does pick out relatively more commercials than news stories, it is not accurate enough in itself to be used during retrieval, and would need to be combined with other cues for more reliable commercial identification. Tags were generated using a threshold of 10dB (NLE=-1.3), but these were not used in the retrieval system for the reason mentioned.

θ	ml	ABC	PRI	CNN	
-1.5	-	36.9@3.2	37.4@15.5	59.2@13.9	
-1.3	-	22.0@1.5	27.6@ 9.5	44.9@ 7.0	
-1.3	20s	9.5@0.2	15.6@ 4.1	23.0@ 1.3	
a) Development data (January TDT-2)					VOA
-1.5	-	39.3@3.4	49.2@26.1	53.0@13.7	18.8@4.8
-1.3	-	23.7@1.7	40.0@17.6	41.5@ 7.2	13.9@2.7
-1.3	20s	8.6@0.2	25.0@ 7.6	21.5@ 1.5	3.7@1.0

b) TREC-9 test data

Table 5: Percentage non-story @ story rejection when using a threshold, θ, on the normalised log energy for 5s windows, including restricting the minimum length, ml.

2.3. Repeat and Commercial

Direct audio repeats (i.e. re-broadcasts) were found using the technique described in [14], by comparing all the audio (across the entire 5 months) from each broadcaster. Commercials were postulated in a similar way to that described in [12], by assuming that segments which had been repeated several times were commercials and that no news portion of less than some smoothing length could exist between them. Table 6 shows the results from applying the parameter set used in the evaluation (C-E) and a less conservative run (C-2) as a contrast. The numbers for our TREC-8 commercial detection system are given for comparison.

	Time (h)		TREC-8	TREC-9	
Br.	N-St	St.	C-E	C-E	C-2
ABC	19.5	42.9	65.5@0.02	79.8@0.01	83.3@0.13
CNN	73.3	170	35.7@0.46	62.4@0.43	69.8@0.62
PRI	11.6	81.5	16.6@0.10	24.5@0.14	28.0@0.19
VOA	9.4	92.9	5.0@0.04	7.2@0.09	8.1@0.11
ALL	114	388	36.3@0.23	57.0@0.24	62.7@0.35

Table 6: Overall time and percentage of non-stories @ stories rejected using both the TREC-8 and TREC-9 commercial detection systems with a less conservative C-2 run for comparison.

Detection performance with this strategy is very impressive, with over half the adverts being identified for negligible loss of news content. Removing these postulated commercials automatically before retrieval was earlier shown not only to reduce the amount of processing necessary but also to significantly improve performance on the TREC-8 data [15]. The improvement from the

[4] NLE is related to the dB from the maximum energy in the episode by:
`ln10 * dB = 10 * ( 1 - NLE )`

TREC-8 to the TREC-9 commercial detection system is due to the change in rules which allows both segments for any given match to be noted within the SDT file[5].

3. TRANSCRIPTIONS

3.1. s1 Transcriptions

The transcriptions used for our s1 runs were those we generated for the 1999 TREC-8 SDR evaluation. A summary of the system is shown in Figure 1 and a detailed description can be found in [12]. The system ran in 13xRT[6] and gave a Word Error Rate (WER) of 15.7% on the November 1998 Hub4 eval data and 20.5% on the 10-hour scoring subset of the TREC-8 data.

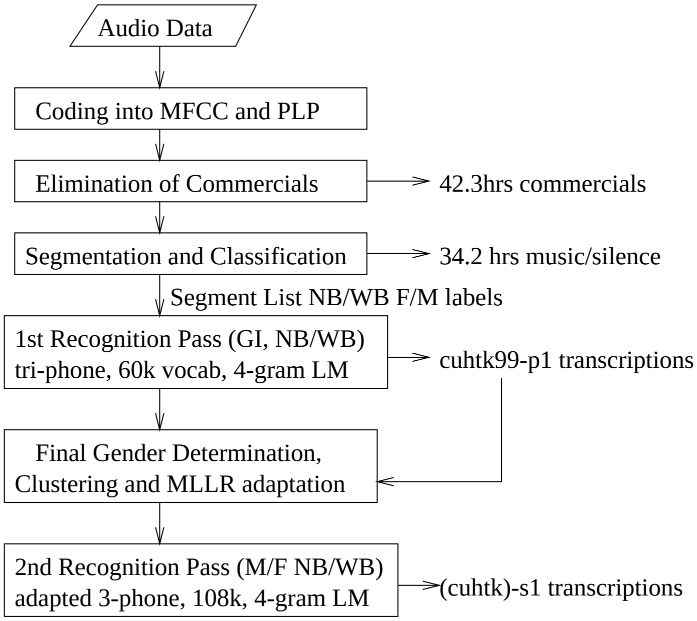
Audio Data
Coding into MFCC and PLP
Elimination of Commercials
42.3hrs commercials
Segmentation and Classification
34.2 hrs music/silence
Segment List NB/WB F/M labels
1st Recognition Pass (GI, NB/WB)
tri-phone, 60k vocab, 4-gram LM
cuhtk99-p1 transcriptions
Final Gender Determination,
Clustering and MLLR adaptation
2nd Recognition Pass (M/F NB/WB)
adapted 3-phone, 108k, 4-gram LM
(cuhtk)-s1 transcriptions

Figure 1: System used to generate transcriptions.

3.2. Other Available Transcriptions

Manually generated closed-caption transcriptions[7] were available for the stories within the SK part of the evaluation from TREC-8 [8]. Word-level time stamps for these portions were produced by LIMSI using forced alignment after some text normalisation. Reference transcriptions were also made for the remaining untranscribed portions of the data by NIST using ROVER on the available TREC-8 ASR transcriptions [7]. The subsequent reference r1 was thus considerably different to the corresponding set of reference transcriptions for TREC-8.

Additional transcriptions were made available for the TREC-9 SDR runs. The baseline cases from TREC-8 SDR produced by NIST using the BBN Rough'N'Ready recogniser [3] were re-released with b1 from TREC-8 becoming cr-nist99b1, whilst b2 from TREC-8 became the baseline b1 for TREC-9. The TREC-8 transcriptions from Sheffield [2] and LIMSI [9] were re-released as cr-shef-s1 and cr-limsi-s1, whilst both sites provided new (higher quality) transcriptions named cr-shef-s2 [2] and cr-limsi-s2 [10] respectively. The WER for these sets of transcriptions on the 10hr TREC-8 scoring subset of the corpus are shown in Table 7.

Recogniser	Corr.	Sub.	Del.	Ins.	WER
r1	91.9	2.5	5.6	2.2	10.3
(cuhtk-)s1	82.4	14.0	3.7	2.9	**20.5**
cr-limsi-s2	82.1	14.2	3.7	3.3	21.2
cr-limsi-s1	82.0	14.6	3.4	3.5	21.5
cr-cuhtk99-p1	77.3	18.5	4.2	3.9	26.6
b1	76.5	17.2	6.2	3.2	26.7
cr-nist99b1	75.8	17.8	6.4	3.3	27.5
cr-shef-s2	74.6	20.0	5.4	3.8	29.2
cr-shef-s1	71.9	22.0	6.1	3.9	32.0

Table 7: WER on TREC-8 10 hour scoring subset of eval. data.

4. SU DEVELOPMENT

4.1. The Basic System

The basic framework for the SU system, shown in Figure 2, is similar to our TREC-8 system [12]; but it does not enforce boundaries at proposed commercial breaks, it uses a different method of performing query expansion and is simpler in not having part-of-speech query weighting, semantic poset indexing or parallel collection frequency weighting.

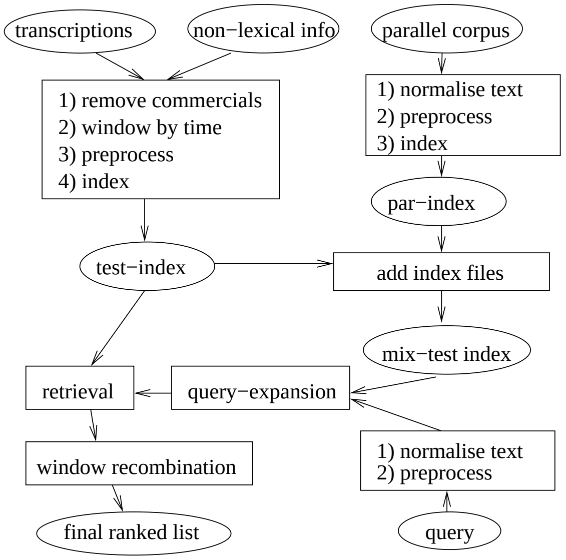

Figure 2: Framework for the SU system.

The transcriptions were first filtered, removing all words which occurred within periods labelled as commercial in the non-lexical file (see section 2.3). Windows of 30s length with an inter-window shift of 15s were then generated to divide up the continuous stream of transcriptions.

[5] In TREC-8, the commercial detection was done pre-recognition in an *on-line* manner i.e. you could not add information about past events retrospectively.

[6] On a Pentium III 550MHz processor running Linux.

[7] Closed-caption transcriptions often use paraphrases or summaries hence giving a significant WER.

Text-normalisation was applied to the query and parallel corpus to minimise the mismatch between the ASR transcriptions and the text-based sources. Preprocessing including mapping phrases and some stemming exceptions, punctuation removal, stop word removal and stemming using Porter's algorithm, for all documents and queries. The stoplist included numbers since some development experiments suggested this increased performance slightly.

The retrieval engine was similar to that employed in TREC-8 [12], using the sum of the combined-weights (CW) [20] for each query term to give the score for any given document. For all runs, the value of K used in the CW formula was 1.4, whilst b was set to 0.6 when story boundary information was present (e.g. when using the parallel corpus) or 0 when no document-length normalisation was necessary (e.g. on the windowed test collection). The inclusion of both query and document expansion before the final retrieval stage is discussed in section 4.2.

The final recombination stage pooled all windows which were retrieved for a given query which originated within 4 minutes of each other in the same episode. Only the highest scoring window was retained, with the others being placed in descending order of score at the *bottom* of the ranked list. Although this means that temporally close stories cannot be distinguished, we assume that the probability that two neighbouring stories are distinct but are both relevant to the same query is less than the probability they are from the same story which drifts in and out of relevance. Although alternative, more conservative strategies are also in use (see e.g. [2]), this strategy proved effective in development experiments [15].

4.2. Document and Query Expansion

4.2.1. Query Expansion

Blind Relevance Feedback (BRF) was used to expand the queries prior to the final retrieval stage within our TREC-8 system [12]. The implementation of query expansion used for TREC-9 differs from this in two main ways. The first concerns which index files to use for the expansion, and the second how to weight the query terms after the expansion stage.

In previous work we ran blind relevance feedback first on the parallel corpus only (PBRF), followed by another run on the test corpus alone (BRF) before the final retrieval stage (e.g. [12]). The idea behind this 'double' expansion was to use the larger parallel corpus, which contained knowledge of story boundaries and had no transcription errors, to add *robustly* related terms to the query before running the standard BRF technique on the test collection. Including both stages of BRF was found to be helpful to performance [16]. However, we have found it very sensitive to the number of terms added, t, and number of documents assumed relevant, r, for each stage. Recent work has used a single stage of query expansion on the union of the parallel and test collections (UBRF) before the final retrieval stage [28]. This gives similar results but is less sensitive to the values of t and r chosen and hence was used in the TREC-9 system.

The method of adding and re-weighting terms during query expansion was changed from TREC-8 to follow the specifications given in [25] and [26] more strictly. All terms were ranked using their Offer Weights (OW), but only those which did not occur in the original query were then considered as potential terms for expansion. The final matching score was obtained by using the MS-RW formula as described on page 798 of [26]. Unlike in previous years, both the original terms and the new expanded terms were reweighted using their Relevance Weight (RW).

4.2.2. Document Expansion

Whilst document expansion has been shown to be beneficial for the case where story boundaries are known [22, 23, 28], it does not seem to have been explored for the SU case. We therefore implemented a document expansion stage for our SU windowing system based on that used in our TREC-8 SK system [12], namely:

1. Form a pseudo-query for each window containing more than 10 different terms, consisting of each distinct term
2. Run this pseudo-query on the parallel collection, giving equal weight to all terms
3. Find the top t expansion terms with the highest Offer Weight from the top r documents
4. Add each expansion term to the window once (i.e. increase the term frequency for each expansion term by 1)

Experiments varying the values of t and r showed that the best performance was obtained for $t = 100, r = 15$ for the TREC-8 queries. This document-expanded index file was then used for the final retrieval stage along with the queries generated *before* document expansion.

4.2.3. Results

The results from including query and document expansion within the SU system on TREC-8 queries are summarised in Table 8 and graphically illustrated in Figures 3 and 4.

When there is no query expansion, document expansion increases mean average precision by 25% and 15% relative for short and terse queries respectively. For moderate query expansion (e.g. $t \leq 8$), document expansion is beneficial for both short and terse queries, but this advantage disappears as the level of query expansion increases. Although the best result for the short queries is obtained when including document expansion (51.72% vs 51.53%), the best performance for the terse queries is considerably worse when including document expansion (47.65% vs 50.56%) and thus it was *not* included in the final system.

The values of $t = 20, r = 26$ were chosen for the UBRF stage despite the fact that they were not optimal for either the short or the terse queries, since they provided more consistent performance across the different query sets.

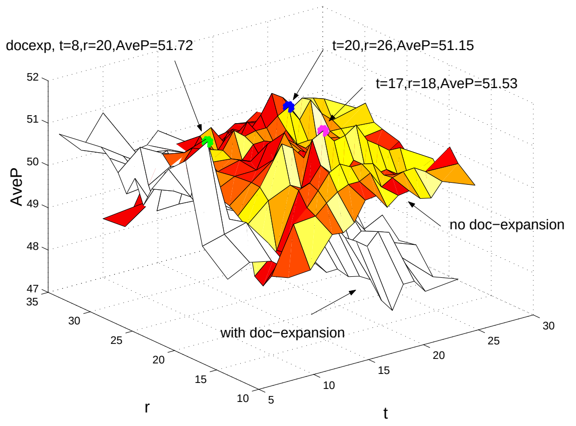

Figure 3: Effect of Query and Document Expansion on TREC-8 *short* queries for SU task on `s1` transcriptions.

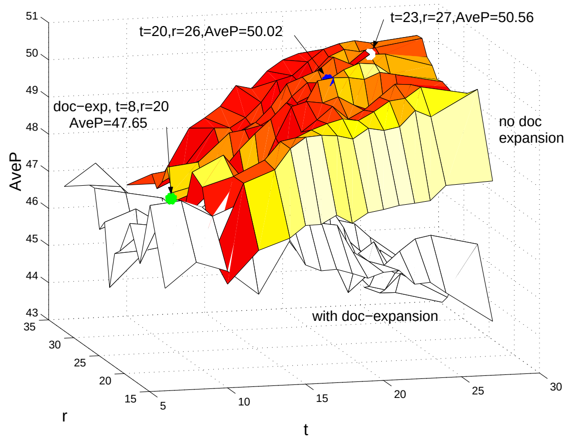

Figure 4: Effect of Query and Document Expansion on TREC-8 *terse* queries for SU task on `s1` transcriptions.

DocExp		QryExp		Short Q		Terse Q	
t	r	t	r	AveP	R-P	AveP	R-P
-	-	-	-	30.89	33.92	32.51	36.77
-	-	8	20	50.84	52.54	47.28	48.43
-	-	17	18	**51.53**	51.78	49.37	49.66
-	-	*20*	*26*	*51.15*	*51.78*	*50.02*	*49.79*
-	-	23	27	51.06	51.60	**50.56**	50.02
100	15	-	-	38.68	42.42	37.27	41.01
100	15	8	20	**51.72**	52.61	**47.65**	48.70
100	15	17	18	49.19	49.17	47.00	47.90
100	15	20	26	48.94	49.27	46.67	49.45
100	15	23	27	49.03	49.45	44.48	47.32

Table 8: Interaction of Query and Document Expansion on SU task on `s1` transcriptions.

4.3. Changing the Window Skip

Recent work at Sheffield [19] suggested that increasing the overlap between windows by decreasing the skip during window generation could help improve performance. A contrast run with their lower skip time was thus made to see if this would have helped our system. The results, given in Table 9, show that this would not have been beneficial to our system, which uses a significantly different method of final window recombination to that used in Sheffield's system.

	Short Queries		Terse Queries	
Windowing System	AveP	R-P	AveP	R-P
length 30s, skip 15s	**51.15**	51.77	**50.02**	49.79
length 30s, skip 9s	48.35	50.27	47.25	48.67

Table 9: Effect of reducing the skip size in window generation for `s1` transcriptions for SU TREC-8 queries.

4.4. Summary

Thus to summarise, after our trials with the TREC-8 queries, our TREC-9 SU evaluation system used windowing, filtering of potential commercials, relatively simple indexing, query but not document expansion, standard Okapi weighting and post-retrieval merging. The query expansion was performed on the union of the test and the parallel text collections.

5. THE FINAL TREC-9 SU SYSTEM

The results using the TREC-9 evaluation SU system on all transcriptions are given in Tables 10 and 11 for the (development) TREC-8 and (evaluation) TREC-9 query sets respectively, whilst the relationship between performance and WER is illustrated in Figure 5.

Transcriptions		Short Q.		Terse Q.	
ID	WER	AveP	R-P	AveP	R-P
r1	10.3	51.04	51.86	48.87	50.77
(cuhtk)-s1	20.5	**51.15**	51.78	**50.02**	49.79
cr-limsi2	21.2	50.90	51.07	49.76	50.03
cr-limsi1	21.5	48.75	49.42	47.47	48.09
cr-cuhtk99p1	26.6	49.34	50.92	47.18	47.88
b1	26.7	48.08	48.92	48.17	48.89
cr-nist99b1	27.5	48.37	49.05	47.86	48.36
cr-shef2	29.2	48.30	50.42	47.69	47.45
cr-shef1	32.0	46.91	48.75	46.55	47.38

Table 10: Cross-recogniser results for (development) TREC-8 queries using the TREC-9 SU evaluation system.

The results confirm the conclusions from earlier work in SDR [8], that the decline in performance as WER increases is fairly gentle (-0.17%AveP/%WER on average here). The relative degradation with WER for the TREC-9 and TREC-8 short queries is almost identical (-0.21 vs -0.20 %AveP/%WER), showing that this fall-off is not query-set specific[8].

[8] TREC-8 terse queries have a slightly different degradation, but were generated in house with different people and restrictions to those for TREC-9.

Transcriptions		Short Q.		Terse Q.	
ID	WER	AveP	R-P	AveP	R-P
r1	10.3	**40.03**	42.09	**44.02**	47.38
(cuhtk)-s1	20.5	38.83	40.36	42.99	45.02
cr-limsi2	21.2	37.24	39.28	41.62	44.12
cr-limsi1	21.5	36.56	38.57	40.19	43.68
cr-cuhtk99p1	26.6	37.26	39.49	40.44	42.92
b1	26.7	37.08	39.91	40.75	43.87
cr-nist99b1	27.5	36.08	39.86	40.99	44.39
cr-shef2	29.2	37.03	39.48	39.83	42.65
cr-shef1	32.0	36.44	38.96	39.58	42.42

Table 11: Cross-recogniser results for the TREC-9 SU eval.

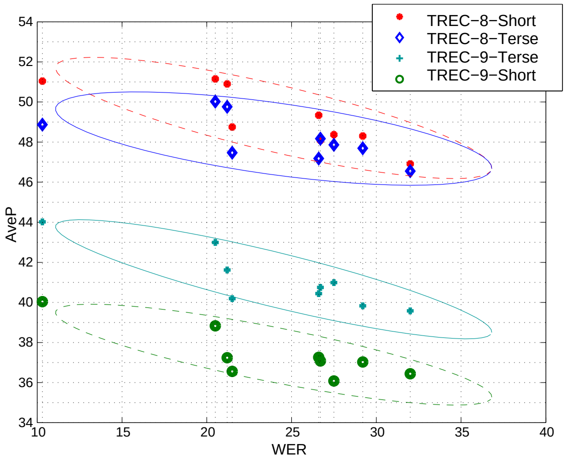

Figure 5: Relationship between WER and AveP for the TREC-9 system on TREC-8 and TREC-9 queries. The ellipses represent 2 standard-deviation points.

The performance on the TREC-8 (development) queries is significantly higher than that on the TREC-9 (evaluation) queries. This may be in part due to three reasons, namely

1. The parameters were tuned for the TREC-8 queries, and may thus be sub-optimal for the TREC-9 queries.

2. All commercials and "filler" portions (e.g. those which summarise stories coming up) were also evaluated for relevance in TREC-9, whereas they were assumed *irrelevant* for TREC-8. Over the 50 TREC-9 queries, there were 93 instances of these portions being scored as relevant. Since our system tries to remove portions such as these by automatically removing commercials before retrieval and biasing the post-processing towards removing fillers[9], the new relevance assessment procedure may have detrimentally affected our score.

3. Natural variation in query difficulty may have meant the TREC-9 queries were "harder" than the TREC-8 ones[10].

To investigate point 2 further, the TREC-9 runs were re-scored using the TREC-8 procedure, which assumed all non-news portions were irrelevant. This increased Average Precision by 1.9% on average for the `b1`, `s1` and `r1` runs for both query sets. This is partly because our SU system tries to filter out the non-news portions before retrieval.

The number of relevant stories from each episode for each query was counted to investigate the validity of the assumption made during post-processing, that the probability of a given episode containing more than one relevant story for a given query was small. The results illustrated in Figure 6 show that 72% of all the relevant stories are unique to their episode and query, but there remains the potential to increase performance by altering the post-processing strategy to allow more temporally close distinct hits[11].

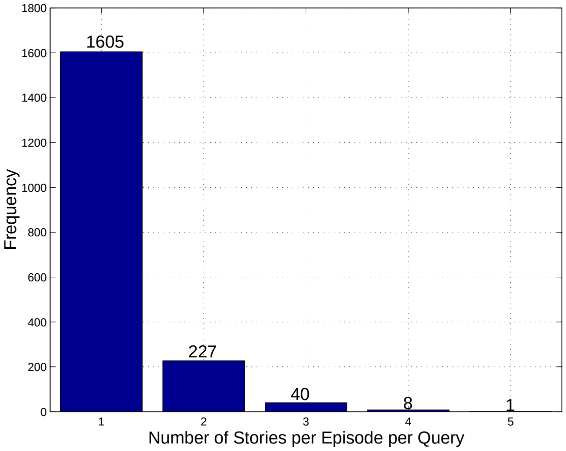

Figure 6: Number of relevant "stories" from each episode for each TREC-9 query .

The expansion parameters were chosen so that the results for the terse and short TREC-8 queries were similar, meaning *sub-optimal values were chosen when considering the short queries alone*. When compared to the (similar) system from Sheffield, whose parameters were chosen based *solely on the short queries*, we do more poorly on average for the short-query runs, but our results are better for all terse query runs [19].

In addition, the parameters $t = 17, r = 18$ which gave the best performance on the short development queries, give better performance on the TREC-9 short queries (AveP=39.38% on `s1`), but worse on the terse queries (AveP=42.78% on `s1`). This suggests that the choice of parameters should take the expected test query length into account and that performance over a wide range of queries might be increased if the expansion parameters were made to be functions of query length.

[9] By finding only the most relevant portions within a short temporal span in each episode.

[10] For the `r1` run, we got <10% AveP for 8 TREC-9 short queries, but only 3 TREC-8 short queries.

[11] For example, q153 has 5 relevant "stories" from the episode 19980528_2000_2100_PRI_TWD, with start times: 235/371/810/1594/1711 seconds, but post-processing merging 4-minute portions means a maximum of 3 could be retrieved using this strategy.

6. THE EFFECTS OF USING NON-LEXICAL INFORMATION

As mentioned in section 1.1, non-lexical information automatically derived from the audio could be used within retrieval in the TREC-9 evaluation. Thus, as discussed in section 2, we generated information for `segment`, `gender`, `bandwidth`, `nospeech`, (high-)`energy`, `repeat` and `commercial` tags directly from the audio.

For the SU system we used the `commercial` tags to filter out words thought to have originated in commercial breaks, but we made no use of the other tags. Thus for our required SN contrast run, we ran the SU system without filtering out the commercials[12]. As can be seen from Table 12, as well as reducing the amount of data processing by around 13%, filtering out commercials improved performance by a small, but statistically significant[13] amount on both sets of development queries across all 3 transcriptions (`r1,s1,b1`). For the TREC-9 evaluation queries only the `s1-terse` and `r1-terse` comparisons were statistically significant[14].

Query	Run	Time	Short Q.		Terse Q.	
Set	ID	Reject.	AveP	R-P	AveP	R-P
TREC-8	SN-r1	0	50.25	50.95	48.18	49.83
	SN-s1	76.2h	50.77	51.20	49.93	50.12
	SN-b1	0	47.86	48.37	47.96	48.85
TREC-8	SU-r1	65.8h	51.04	51.86	48.87	50.77
	SU-s1	92.5h	51.15	51.78	50.02	49.79
	SU-b1	65.8h	48.08	48.92	48.17	48.89
TREC-9	SN-r1	0	40.54	42.50	44.75	47.03
	SN-s1	76.2h	39.00	40.35	42.65	45.11
	SN-b1	0	37.81	40.44	42.17	44.77
TREC-9	SU-r1	65.8h	40.03	42.09	44.02	47.38
	SU-s1	92.5h	38.83	40.36	42.99	45.02
	SU-b1	65.8h	37.08	39.91	40.75	43.87

Table 12: Effect of automatically removing commercials (SU).

Contrast runs were also performed on the development queries using the less conservative `comm2` system and the manual boundaries derived from the SK case. As can be seen from Table 13 using either of these would have resulted in little difference in performance for our own transcriptions. (none significant at the 2% level.)

Other experiments were run for fun on the TREC-8 queries to see the effect of removing various parts of the audio using the non-lexical information, such as high-energy regions, or particular bandwidth/gender segments. The results are given in Table 13 for the `s1` transcriptions, and plotted in Figure 7. The trend is roughly linear, with the best AveP to time-retained ratio being 0.163%AveP/hr when removing all male speakers, whilst the worst is 0.120%AveP/hr when removing female speakers.

Transcriptions			Short Q.		Terse Q.	
Comm.	Reject	ID	AveP	R-P	AveP	R-P
TREC-9	72.8h	r1	50.82	51.69	48.64	51.08
comm2	96.4h	s1	50.72	51.91	49.79	49.59
	72.8h	b1	48.21	49.25	48.31	49.07
manual	113.9h	r1	51.18	52.75	49.37	51.46
comms	126.6h	s1	50.97	52.07	50.18	50.33
(ndx file)	113.9h	b1	48.28	48.66	49.16	49.71
no loud	111.2h	s1	47.92	49.60	47.29	47.93
no nb	127.4h	s1	46.39	48.52	45.56	46.20
no wb	450.1h	s1	7.69	11.26	8.08	11.29
no male	347.9h	s1	25.25	30.02	25.28	30.64
no female	229.6h	s1	32.59	37.33	31.74	36.69

Table 13: Effect of including non-lexical information for TREC-8 queries. (`s1` reject times include time removed in TREC-8 commercial detection and segmentation stages.)

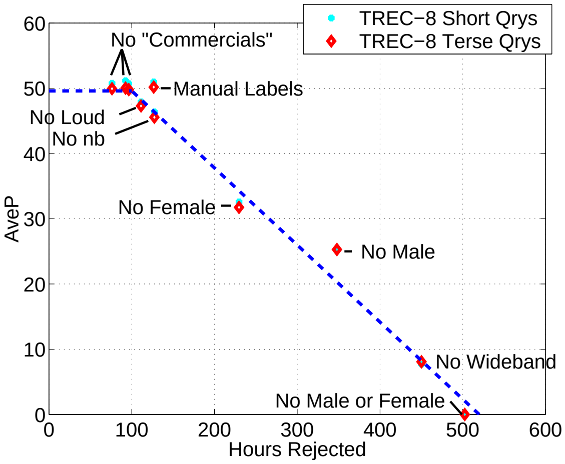

Figure 7: Effect of removing data using non-lexical information on TREC-8 queries for `s1` transcriptions.

7. THE STORY-KNOWN (SK) CONTRAST RUN

The SK system was similar to the SU system described in section 4. The commercial-removal, window-generation and post-merging stages were no longer necessary, since the known story boundaries defined the documents in the collection, but the rest of the system remained practically unaltered.

Document expansion was performed in the same way as described in section 4.2.2 except that the pseudo-query for each document was defined as the 100 terms from the document with the lowest collection frequency. Different values of t and r were investigated for the document expansion stage, but there proved to be little difference between the results, so the values of $t = 200, r = 10$ were chosen to be compatible with [28].

UBRF was performed as described in section 4.2.1, using the *un-expanded* document file to expand the query which was then

[12] Note that the `s1` transcriptions already had 76.2hrs of audio filtered out from the TREC-8 segmentation and commercial detection stages [12].

[13] Using the Wilcoxon Matched-Pair Signed-Rank test at the 5% level. (see [16] for discussion of the usage of this test.)

[14] Using the TREC-8 scoring procedure, (non-news portions are assumed irrelevant), *all* TREC-9 SU runs performed better than the corresponding SN runs.

run on the *expanded* document file, and the values of $b = 0.6, k = 1.4$ were retained for all retrieval stages. Results for varying the expansion parameters in the UBRF stage for the SK system are illustrated in Figures 8 and 9 for the short and terse TREC-8 queries and are summarised in Table 14.

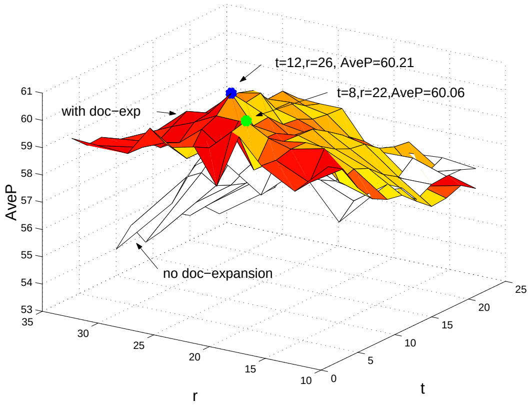

Figure 8: Effect of Query and Document Expansion on TREC-8 *short* queries, SK case, `s1` transcriptions.

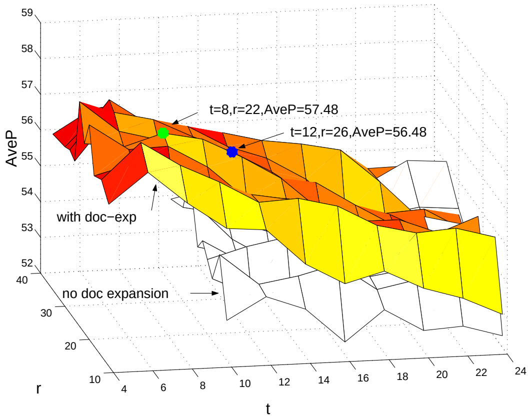

Figure 9: Effect of Query and Document Expansion on TREC-8 *terse* queries, SK case, `s1` transcriptions.

The inclusion of document expansion improved performance across both development query sets and all 3 transcriptions, with the largest improvements when the level of query expansion was low to moderate. This consistent improvement was not found for the SU case. The difference is thought to be because the pseudo-queries from windowing for the SU case may be multi-topic, and cannot be as long as for the SK case, since the windows must be kept small (e.g. around 30s) to obtain acceptable performance.

The values of $t = 8, r = 22$ were chosen for the UBRF stage for the SK run to give good performance across both development query sets when used in conjunction with document expansion. The amount of query expansion for the SK case was thus chosen to be less than that used for the SU case because of the interaction between the query and document expansion devices.

The SK results on the TREC-9 evaluation queries are given in Table 15. Since this used a subset of the data and hence also a different relevance file to the SU case, another SK run across *all* the data was performed to allow a more direct comparison between SK and SU cases.

	DocExp		QryExp		Short Q		Terse Q	
Tr.	t	r	t	r	AveP	R-P	AveP	R-P
s1	-	-	-	-	46.29	45.85	45.67	44.53
s1	-	-	8	22	57.41	55.89	54.31	51.31
s1	-	-	12	26	59.11	57.14	54.04	50.65
s1	200	10	-	-	50.76	49.42	52.91	51.67
s1	*200*	*10*	*8*	*22*	*60.06*	*57.62*	***57.48***	*55.15*
s1	200	10	12	26	**60.21**	56.84	56.48	54.88
r1	-	-	-	-	48.19	47.69	47.44	46.28
r1	-	-	8	22	58.17	57.73	54.63	53.19
r1	200	10	-	-	51.65	52.27	53.65	53.76
r1	*200*	*10*	*8*	*22*	*59.04*	*57.31*	*56.95*	*56.20*
b1	-	-	-	-	43.31	43.32	43.17	41.86
b1	-	-	8	22	55.19	54.10	53.04	50.52
b1	200	10	-	-	49.56	48.94	50.86	49.46
b1	*200*	*10*	*8*	*22*	*58.18*	*55.69*	*55.88*	*54.20*

Table 14: Interaction of Query and Document Expansion on SK task for TREC-8 queries.

	Short Queries			Terse Queries		
	SK	SK	SU	SK	SK	SU
ID	AveP	R-P	AveP	AveP	R-P	AveP
r1(a)	49.60	47.05	—	52.68	49.26	—
s1(a)	49.47	47.83	—	51.94	50.26	—
b1(a)	48.31	47.38	—	50.44	48.85	—
r1(b)	47.44	45.74	40.04	50.99	48.20	44.02
s1(b)	46.42	44.93	38.83	49.18	48.40	42.99
b1(b)	46.55	46.52	37.08	48.56	47.62	40.75

Table 15: Comparison of TREC-9 SK and SU results. (a) is on the 21,754 story subset, whilst (b) is on all the data, to allow a fairer comparison with the SU case.

Although our SU-SDR system has been improved by around 20% relative[15] since the TREC-8 evaluation [12], and the gap between SK and SU has been reduced from 14% AveP to 8%, there still remains a considerable performance gap between the SK and SU cases.

8. CONCLUSIONS

This paper has described work carried out at Cambridge University for the TREC-9 SDR evaluation. The experiments confirmed that the relative degradation of Average Precision with increasing recogniser error rate is gentle, and performance on high-quality ASR transcriptions can be as good as that on a manually transcribed reference.

Standard indexing techniques and Okapi-weighting provide a good baseline system and adding query expansion using the union

[15] Comparing AveP for `s1` on TREC-8 short queries

of the test and a contemporaneous parallel newswire collection increases performance further. Including a windowing and post-retrieval recombination strategy allows good performance even when no story boundaries are known in advance. Document expansion, which previously has been found to work well for the SK case, was extended to the SU framework and shown to improve performance for small to moderate levels of query expansion.

Non-lexical information derived directly from the audio, which would not normally be transcribed, can be used to improve real SDR systems. Audio repeats can accurately predict the presence of commercials, which can be filtered out before retrieval, and some broadcast structure information can be recovered by analysing cues such as bandwidth, signal energy and the presence of music in the audio. Browsing and understanding could also be improved by including tags such as sentence boundaries and speaker turns. Optimally integrating non-lexical information within real SDR systems, using larger databases and including other information such as video data provide interesting challenges for the future.

Acknowledgements

This work is in part funded by an EPSRC grant reference GR/L49611.

9. REFERENCES

[1] D. Abberley, S. Renals, G. Cook & T. Robinson *Retrieval of Broadcast News Documents with the THISL System*. Proc. TREC-7, pp. 181-190, 1999

[2] D. Abberley, S. Renals, D. Ellis & T. Robinson *The THISL SDR System*. Proc. TREC-8, pp. 699-706, 2000

[3] C. Auzanne, J.S. Garofolo, J.G. Fiscus & W.M Fisher *Automatic Language Model Adaptation for Spoken Document Retrieval*. Proc. RIAO 2000, Content-Based Multimedia Information Access, pp. 132-141, 2000

[4] R. Ekkelenkamp, W. Kraaij & D. van Leeuwen *TNO TREC7 site reports: SDR and filtering*. Proc. TREC-7, pp. 519-526, 1999

[5] M. Franz, J.S. McCarley, R.T. Ward *Ad hoc, Cross-language and Spoken Document Information Retrieval at IBM*. Proc. TREC-8, pp. 391-398, 2000

[6] J.S. Garofolo, J. Lard, C.G.P. Auzanne & E.M. Voorhees 2000 TREC-9 Spoken Document Retrieval (SDR) Track Evaluation Specification.
http://www.nist.gov/speech/tests/sdr/sdr2000/sdr2000.htm

[7] J.S. Garofolo, J. Lard & E.M. Voorhees 2000 TREC-9 Spoken Document Retrieval Track: Overview and Results. To appear in Proc. TREC-9

[8] J.S. Garofolo, C.G.P. Auzanne & E.M. Voorhees *The TREC Spoken Document Retrieval Track: A Success Story*. Proc. RIAO 2000, Content-Based Multimedia Information Access, pp. 1-20, 2000

[9] J.-L. Gauvain, Y. de Kercadio, L. Lamel & G. Adda *The LIMSI SDR System for TREC-8*. Proc. TREC-8, pp. 475-482, 2000

[10] J.-L. Gauvain, L. Lamel, C. Barras, G. Adda & Y. de Kercadio *The LIMSI SDR System for TREC-9*. To appear in Proc. TREC-9

[11] T. Hain, S.E. Johnson, A. Tuerk, P.C. Woodland & S.J. Young *Segment Generation and Clustering in the HTK Broadcast News Transcription System* Proc. DARPA Broadcast News Transcription and Understanding Workshop, pp. 133-137, 1998

[12] S.E. Johnson , P. Jourlin, G.L.Moore, K. Spärck Jones & P.C. Woodland *Spoken Document Retrieval for TREC-8 at Cambridge University*. Proc. TREC-8, pp. 197-206, 2000

[13] S.E. Johnson *Who Spoke When? - Automatic Segmentation and Clustering for Determining Speaker Turns*. Proc. Eurospeech, Vol. 5, pp. 2211-2214, 1999

[14] S.E. Johnson, P.C. Woodland *A Method for Direct Audio Search with Applications to Indexing and Retrieval*. Proc. ICASSP 2000,Vol. 3, pp. 1427-1430, 2000

[15] S.E. Johnson , P. Jourlin, G.L.Moore, K. Spärck Jones & P.C. Woodland *Audio Indexing and Retrieval of Complete Broadcast News Shows*. Proc. RIAO 2000, Content-Based Multimedia Information Access, Vol. 2, pp. 1163-1177, 2000

[16] P. Jourlin, S.E. Johnson , K. Spärck Jones & P.C. Woodland *Spoken Document Representations for Probabilistic Retrieval*. Speech Communication, Vol 32, No. 1-2, Sept. 2000, pp. 21-36

[17] C. Ng, R. Wilkinson & J. Zobel *Experiments in spoken document retrieval using phoneme n-grams* . Speech Communication, Vol 32, No. 1-2, Sept. 2000, pp. 61-77

[18] D. Oard *User Interface Design for Speech-Based Retrieval*. Bulletin of the American Society for Information Science, 26(5) pp. 20-22, June/July 2000

[19] S. Renals & D. Abberley *The THISL SDR system at TREC-9*. To appear in Proc. TREC-9

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

[21] E. Scheirer & M. Slaney *Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator*. Proc. ICASSP'97, pp. 1331-1334, 1997

[22] A. Singhal, J. Choi, D. Hindle & D.D. Lewis *AT&T at TREC-7*. Proc. TREC-7, pp. 239-251, 1999

[23] A. Singhal & F. Pereira *Document Expansion for Speech Retrieval*. Proc. SIGIR '99, pp. 34-41, 1999

[24] A. Singhal, S. Abney, M. Bacchiani, M. Collins, D. Hindle & F. Pereira *AT&T at TREC-8*. Proc. TREC-8, pp. 317-330, 2000

[25] K. Spärck Jones, S. Walker, S.E. Robertson *A probabilistic model of information retrieval : Development and status*. Technical report, TR-446, Computer Laboratory, University of Cambridge, 1998.

[26] K. Spärck Jones, S. Walker, S.E. Robertson *A probabilistic model of information retrieval : Development and comparative experiments, Parts 1 and 2*. Information Processing and Management 36(6) pp. 779-840, 2000

[27] A. Tuerk, S.E. Johnson , P. Jourlin, K. Spärck Jones & P.C. Woodland *The Cambridge University Multimedia Document Retrieval Demo System*. Proc. RIAO 2000, Content-Based Multimedia Information Access, Vol. 3, (Applications) pp. 14-15, 2000.

[28] P.C. Woodland, S.E. Johnson, P. Jourlin, & K. Spärck Jones *Effects of Out of Vocabulary Words in Spoken Document Retrieval*. Proc. SIGIR'2000 pp. 372-374, 2000

For TREC publications, see `http://trec.nist.gov/pubs.html`
For Cambridge University SDR papers, see
`http://svr-www.eng.cam.ac.uk/research/projects/mdr/`