Indexing and Retrieval of Broadcast News

Steve Renals(1), Dave Abberley(1), David Kirby(2) and Tony Robinson(3,4)
(1) Department of Computer Science, University of Sheffield, UK
(2) BBC Research and Development, UK
(3) SoftSound Ltd., UK
(4) Department of Engineering, University of Cambridge, UK

8 March 2000

Abstract

This paper describes a spoken document retrieval (SDR) system for British and North American Broadcast News. The system is based on a connectionist large vocabulary speech recognizer and a probabilistic information retrieval system. We discuss the development of a realtime Broadcast News speech recognizer, and its integration into an SDR system. Two advances were made for this task: automatic segmentation and statistical query expansion using a secondary corpus. Precision and recall results using the Text Retrieval Conference (TREC) SDR evaluation infrastructure are reported throughout the paper, and we discuss the application of these developments to a large scale SDR task based on an archive of British English broadcast news.

1 Introduction

Retrieval of audio segments according to their content is a challenging and significant problem. It has been estimated that a large proportion of human-generated information is spoken and that much of this is in the form of television and radio broadcasts (Morrison and Morrison 1998). While the navigation and retrieval of textual data is commonplace, it still an outstanding research problem to perform such operations on archives of spoken data. *Spoken document retrieval* (SDR)—the task of finding those segments of an audio archive that correspond to a user's information need—is one such task. The problem of broadcast news retrieval, in particular, has received considerable attention, not least because of the availability of acoustic training data, and the mix of acoustic conditions that characterize news broadcasts, including substantial sections of planned, noise-free speech.

Two principal approaches have been used for SDR: *phone-based* approaches in which the audio data is transcribed at a phone level and *word-based* approaches in which a large vocabulary speech recognizer is used to transcribe the audio data at the word level. Since queries may be assumed to be at the word level,[1] additional processing is required to build a suitable index from a phone-level transcription: several researchers (Ng and Zue 1998; Smeaton et al. 1998; Kraaij et al. 1998) have

[1] This may not be the case for spoken queries.

used phone-level n-grams; a more elaborate adaptive scheme was employed by Wechsler et al. (1998). Phone-level transcription usually has a high error rate—even on databases such as TIMIT, state-of-the-art recognizers return phone error rates of over 25% (Robinson 1994). To make phone-level approaches more robust, algorithms to scan phone lattices for keywords have been developed (James and Young 1994; Foote et al. 1997; Ferrieux and Peillon 1999). Such lattice approaches are more suitable for applications with a fixed query (eg, filtering or routing), as each new query word demands that the entire archive of phone lattices is scanned—an operation which scales linearly with archive size.[2]

The word-based approach to SDR, using a large vocabulary speech recognizer, was first applied by the Informedia group at Carnegie Mellon University (Hauptmann and Witbrock 1997), and has been adopted by several groups since then (Abberley et al. 1998; Allan et al. 1998; Johnson et al. 1999; Singhal and Pereira 1999), particularly in the framework of the TREC (Text Retrieval Conference)[3] SDR evaluations (Garofolo et al. 1999). These approaches—and the one described in this paper—all use a similar methodology. A large vocabulary speech recognizer is used to provide a word-level transcription; the transcribed audio segment is then treated as a text document by an information retrieval (IR) system.

The advantages of the word-based approach are clear. IR is more robust when applied to words compared with phone n-grams, particular at the high error rates observed when recognizing broadcast data (although explicit modelling of phone recognition errors has been investigated by Wechsler et al. (1998) and Ng and Zobel (1998)). Furthermore, word-level recognition enables the constraints of the pronunciation dictionary and language model to be applied. Two possible disadvantages of the word-based approach, compared with the phone-based approach, are an increased computational burden and a closed vocabulary. In section 3 we show that large vocabulary continuous speech recognition of broadcast speech demands around three times the computation compared with phone recognition, and can be achieved in real-time on a modern PC. Most large vocabulary systems used for SDR have a vocabulary of around 60 000 words;[4] for English language broadcast news speech, this typically corresponds to an out-of-vocabulary (OOV) rate of 1–3%. Such a level of OOV has not yet presented a significant barrier to SDR (indeed, in the last five TREC "ad hoc" evaluations, only 9 out of 900 query words would have been OOV with respect to 65 532 word vocabulary we use for the North American Broadcast News speech recognizer discussed in this paper.[5]

In this paper we discuss some experiments that we have carried out to develop a spoken document retrieval system for British and American English broadcast news, with a target application of developing a system to index the news output of the British Broadcasting Corporation (BBC). This paper summarizes the principal results presented in some earlier conference papers (Abberley et al. 1998; Abberley et al. 1999; Robinson et al. 1999) and includes further work on query expansion algorithms and

[2] An algorithm that lessens this computational burden was introduced by Dharanipragada and Roukos (1998).

[3] http://trec.nist.gov

[4] Much larger vocabularies have been used (eg, Singhal et al. 1999).

[5] The OOV problem is more substantial in languages with compound words, such as German, where the OOV rate on a similar broadcast speech task is typically 3–4 times higher.

automatic segmentation. Our work has focussed on British English, with an archive constructed from around 2.5 hours of BBC news recordings per day. Additionally, we have constructed a North American English system, using the resources gathered for the "Hub-4" broadcast speech recognition evaluations[6] and the TREC SDR evaluations. Access to this evaluation infrastructure has enabled us to evaluate algorithmic developments without having to develop a parallel evaluation framework for British English broadcast news.

The paper is organized as follows. Section 2 outlines the collection of application-specific acoustic and textual data for British English broadcast speech. The large vocabulary speech recognition system is described in section 3, with particular reference to the computationally efficient algorithms employed, the models required for British English broadcast news and evaluation of the speech recognition performance. Section 4 describes the basic IR methods that we have used and section 5 describes the evaluation metrics employed. Two advances have been made for the SDR task: section 6 describes the use of automatic segmentation methods and section 7 describes a query expansion methodology. Throughout we report precision and recall results on the TREC-7 SDR evaluation data; section 10 describes the application of all the reported developments to the British English broadcast news retrieval task.

2 Data Collection

The Hub-4E acoustic and text data, available from the Linguistic Data Consortium[7], was used to train the North American Broadcast News speech recognition system. A similarly controlled and annotated data resource was not available for British English broadcast speech. 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 about 50 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 (eg, at the end of each sentence or speaker turn) as this was particularly labour intensive. The timing of major changes in acoustic condition were noted; otherwise synchronization marks were added every five minutes. We further developed our speech alignment software to take account of the coarse timing information when providing 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. Further text data, totalling around 4 million words, was obtained from British English newspaper and newswire sources.

[6] http://www.itl.nist.gov/iaui/894.01/

[7] http://www.ldc.upenn.edu

3 Broadcast Speech Recognition

For this application we required a speech recognition system that maximized recognition accuracy while meeting practical constraints on decoding time and smooth integration with the rest of the system. Currently the system decodes and indexes several hours of broadcast news per day; it is easy to envisage applications that need to perform continuous, 24 hours/day, recognition and indexing of broadcast material, perhaps across several channels. Therefore we chose to operate in real-time on commonly available computers (Intel 550 MHz PCs). We have used the ABBOT connectionist speech recognition system (Robinson et al. 1996) using a perceptual linear prediction front end and an acoustic model based on two recurrent networks trained on forward-in-time and backward-in-time data. This acoustic model is relatively simple (only context-independent phone models are used) and results in a fast and efficient system that provides a favourable framework for other developments such as confidence measures and pronunciation learning (Robinson, Cook, Ellis, Fosler-Lussier, Renals, and Williams 2000)

Acoustic and language models for the North American broadcast news system are described by Cook et al. (1999). For the British English system we used the corpus described in section 2, with the addition of 130 million words of Hub-4E text data (largely newswire and newspaper text).

In order to achieve real-time recognition we developed a new search algorithm based on a stack decoder. The essence of this algorithm is the reordering of the computations required to perform Viterbi decoding such that the inner loop is over the time index. We have found that this yields considerable time and memory savings, thus enabling the baseline system to run in real-time on a 550MHz Pentium-III using less than 256 Mb RAM, most of which is used to store the language model. Further details of this search technique can be found in Robinson and Christie (1998) and Robinson, Christie, and Cook (2000). In addition, we perform:

Whole show decoding The efficient memory usage of the time-first decoder allows decoding of hour-long shows and so enables the use of online acoustic normalization as an alternative to the more common segment-based normalization 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 realization of a short period of silence allows the decoder to hypothesize sentence boundaries, and so fit the desired functionality of multiple sentence decoding.

For this task we have observed that approximately one third of the recognition time is taken up with computing the acoustic features and evaluating the acoustic models, with the remaining two thirds being used for large vocabulary search. The search time for phone recognition is negligible in this context, hence the difference in overall time between performing phone recognition and large vocabulary recognition is a factor of three. Given that real-time operation is feasible, we believe that the benefits of the word and language model constraints and the convenience of indexing the recognition output at the word level far outweigh the disadvantage of slower recognition. The issue of a finite lexicon is discussed in section 9.

System	WER
baseline system (real-time)	29.2%
7x real-time	28.9%
with North American LM	30.7%
without cross-sentence	29.4%

Table 1: Overall word error rates by ASR variation for a three hour evaluation set of British English radio and television news broadcasts. The baseline system corresponds to real time on a 550MHz Pentium-III.

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. BBC1 is broadcast television news.

Our primary objective is fast, efficient information retrieval. Since speech recognition performance is weakly correlated with this goal, in many cases we are prepared to accept an increase in word error rate (WER) in order to maximize 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, using the language model described above with cross sentence decoding, and online acoustic normalization instead of segmentation-based normalization. The baseline WER is higher than that reported for North American broadcast news (Cook et al. 1999), in part because we decode complete broadcasts and also because we score against single hypothesis transcriptions with no flexibility for reasonable variants. For comparison, a system was built that ran seven times slower (7x real-time), and the associated WER shows that we make only a few more errors in order to run at the speed we desire. It is not expected that any of these error rate changes would have a significant effect on IR 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 recognize. Figure 1 plots the error rate throughout a show measured using a 15 second rectangular window. The dashed lines mark the story boundaries; note that as new topics are introduced by the newsreader a lowering of WER is often observed. There is a large variation in WER within a topic. This has implications for unsegmented IR and related areas such as audio summarization where a concentration on the sections where the speech recognition system makes fewer errors is desirable.

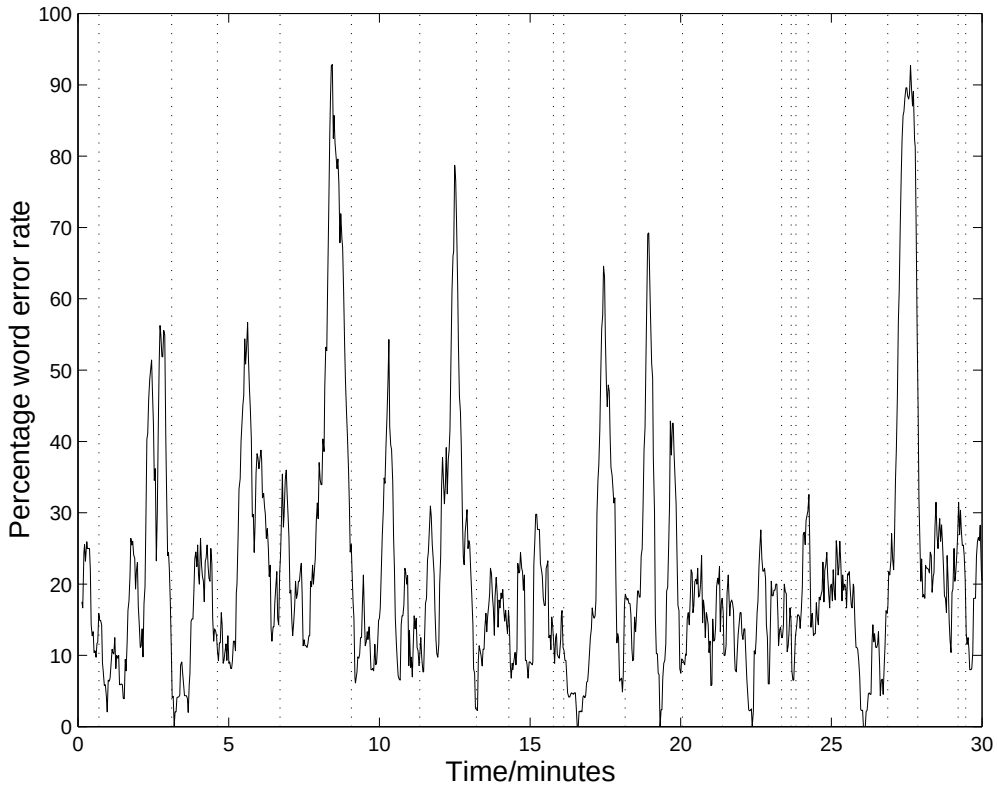

Figure 1: Word error rate over time for Radio 4 News of 10 Feb 1999. Word error rate was computed using a 15s rectangular window; the vertical dotted lines correspond to manually tagged story boundaries.

4 Information Retrieval

The principal goal of an IR system is to return those *documents* that are best matched to a user's *query*. Historically, documents have referred to textual objects; however, we view a document more generally, as a "natural" unit for the situation in question. In the case of spoken document retrieval, a document refers to the audio segment for a particular story (obtaining these segments is discussed in section 6). Since a word-based approach is adopted here, an audio segment may be represented textually, and well investigated methods for IR can be directly applied.

A *ranked* IR system returns documents in order of their relevance to a query. Such a system may be viewed as consisting of three components: text normalization, indexing and matching. Text normalization includes the usual operations of processing any markup and removing punctuation, to result in a stream of words, which is further processed via the operations of *stopping* and *stemming*. Stopping involves deleting those words in the query or document which are viewed as playing no semantic role (van Rijsbergen 1979). Such words include function words and words that commonly prefix queries (eg, "what", "find"). Stemming involves mapping related words to a common root form by automatically stripping off affixes. Although morphological decomposition methods may be used for this operation, consistently good results have been obtained by using simple, rule-based approaches, such as the commonly used Porter stemming algorithm for suffix stripping (Porter 1980). After these operations, we may regard the document as having been processed into a stream of indexing *terms*.

Efficient algorithms and data structures may be used to index a document collection, to enable rapid retrieval of documents containing query words. The key data structure for this operation is an *inverted* or *index* file; computationally efficient compression schemes, such as variable-byte coding of integers, may be used to compress inverted files while still enabling random access (Williams and Zobel 1999).

The operations of text normalization and indexing are essential components of any IR system. Ranked IR systems return a relevance score for each document under consideration with respect to the query. Relevance scores are typically based on a probabilistic model with the assumption that the terms in a document are conditionally independent of each other given the relevance or non-relevance of that document to the query (Robertson and Spärck Jones 1976). Given this independence assumption, a *term-weighting* function which supplies a weight for a term t given a document d may be defined. Such a function usually involves the within-document *term frequency* and the *collection frequency weight*. The term frequency, $TF(t,d)$, is the number of occurences of term t in document d. The collection frequency weight of term t, $CFW(t)$, is a measure of the proportion of the collection in which the term appears:

$$CFW(t) = \log\left(\frac{N}{n(t)}\right), \tag{1}$$

for a collection of N documents, in which term t occurs in $n(t)$ documents. The collection frequency weight is closely related to the inverse document frequency, $IDF(t)$. It can be seen to arise when a binomial is used to model term occurance in relevant and non-relevant documents (Croft and Harper 1979); it may also be interpreted as an estimate of the mutual information between a term and the set of documents (Siegler and Witbrock 1999).

The product of the collection frequency weight and the term frequency results in a straightforward term-weighting function, often referred to as $tf \cdot idf$. Using some further refinements of the probability model (Robertson and Walker 1994; Spärck Jones et al. 1998), a discounted version of $tf \cdot idf$, referred to as the combined weight $CW(t,d)$, may be defined:

$$CW(t,d) = \frac{(K+1) \cdot CFW(t) \cdot TF(t,d)}{K + TF(t,d)}. \tag{2}$$

K may be viewed as a discounting parameter on the term frequency: when K is 0, the combined weight reduces to the collection frequency weight; as K increases the combined weight asymptotically approaches $tf \cdot idf$.

The combined weight (2) may be adjusted to take document length into account (Robertson et al. 1995):

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

where $NDL(d)$ is the length of d normalized by the mean document length across the collection and b is an empirically determined constant that controls the influence of document length ($0 \leq b \leq 1$). When $b = 0$ document length normalization is not applied; as b is increased, the effect of term frequency is reduced (corresponding to an assumption that increased document length is due to repetition). For manually segmented broadcast news transcriptions (and also newswire and newspaper text) an empirically determined value for b of around 0.7 is typically used.

The overall weight of a document d relative to query Q, $W(Q,d)$, is computed by summing the combined weights of each term in the query, relative to the document:

$$W(Q,d) = \sum_{t \in Q} CW(t,d) \tag{4}$$

5 Evaluation

The objective of a ranked IR system is to order a document collection by the probability of relevance to the query. Evaluation may be carried out in terms of precision and recall, relative to the *relevance assessments* made by the user who submitted the query. A recall-precision curve can be generated, which may be summarized into a single *average precision* value which is approximately the area under this curve. Precision values (at a particular ranking threshold) are straightforward to compute; indeed, an alternative measure of performance is the Precision at N, $Prec(N)$, which is the precision obtained for the top N documents, of which N_{rel} are judged as relevant:

$$Prec(N) = \frac{N_{rel}}{N} \tag{5}$$

Recall is less straightforward to estimate than precision, since an exact estimate requires a relevance assessment for each document in the archive, relative to each query. This may be approximated by estimating the set of documents that are potentially relevant; in the TREC evaluations this set is estimated by pooling the outputs of the several

independent IR systems that participate in the evaluation (Harman 1996). Despite this approximation, several thousand relevance assessments per query are often required to estimate the recall.

The spoken document retrieval (SDR) track of TREC has established a pooled relevance assessment infrastructure for the indexing and retrieval of broadcast news. We have taken advantage of this, and our principal set of objective evaluations are within the framework of the TREC-7 SDR track (Garofolo et al. 1999). This consisted of an archive 100 hours of North American broadcast news, with a further 100 hours available for acoustic model training. A set of 23 queries, together with relevance assessments is available for this data.[8] Such an evaluation, which relies on recall measures, would be extremely labour intensive to perform for the BBC data. Hence, our objective evaluations for the British English SDR system use the Precision at N measure.

6 Segmentation

Speech rarely arrives with marked segment boundaries. Although controlled evaluations, such as TREC SDR, have included hand segmentation of news broadcasts into stories, this feature is typically not available for most applications. The corpus we have collected for the BBC application is recorded off air, and some segmentation is necessary to develop an SDR system.

There has been a substantial amount of work in automatically segmenting documents for text retrieval. Callan (1994) and Kaszkiel and Zobel (1997) have investigated so-called *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 (Hearst 1997; Yamron et al. 1998). The benefits of the passage-based approach 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 substantial 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.

To enable the objective evaluation of different automatic segmentations, we have used the TREC-7 SDR corpus since relevance judgments are available. As this is a segmented corpus, some adaptations were necessary to enable automatic segmentation experiments. To simulate the unsegmented condition all segmented stories were abutted and segment boundaries removed. This has the side effect of removing the "gaps"

[8] This is a small archive with a small set of assessed queries. Larger scale SDR experiments would, of course, be desirable. However, at the time of writing this dataset represented the largest scale SDR evaluation data available.

due to unrecognized material such as adverts and sports news. Automatically segmented documents may be characterized by a time index (eg, the segment mid-point); to enable the TREC relevance judgments to be used, these time indexes are converted to the original document IDs at evaluation time.

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 2 shows the average precision for varying window lengths and overlaps, using rectangular windows based on fixed time intervals (top) and fixed word lengths (bottom). 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 seems much smoother for the time-based window.

A side-effect of the automatic segmentation scheme is that adjacent overlapping segments that are part of the same story are likely to produce similar scores. Consequently, the list of retrieved documents will contain many segments from the same news item. To reduce this duplication problem, any overlapping segments occurring in the list of retrieved stories may be combined into a rescored composite story. The problem of how to rescore the combined documents has been investigated experimentally. Several schemes were tried including using the maximum score from the set of documents to be combined, reestimating equation (2) for the combined document (updating CFW, but not accounting for the overlap between adjacent documents), and other, more ad-hoc, methods. The best performing rescoring formula proved to be:

$$W_{\text{DERB}}(Q,D) = \frac{\sum_{s \in D} W(Q,s)}{1 + (S-1)\frac{segskip}{seglen}} \tag{6}$$

where $W_{\text{DERB}}(Q,D)$ is the retrieval score for combined document D (made up of S segment documents s) with respect to query Q. $seglen$ and $segskip$ are the segment window length and segment skip respectively. This formula has the effect of boosting the score of a combined document relative to that of a stand-alone 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.[9]

7 Query Expansion Algorithms

If a relevant document does not contain any of the query terms, then the overall query/document weight (computed using (2) or (3)) 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 OOV query words. *Query expansion* (QE) 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.

[9]Later experiments have indicated that using the maximum score from the set of documents to be combined produced a more consistent improvement in average precision.

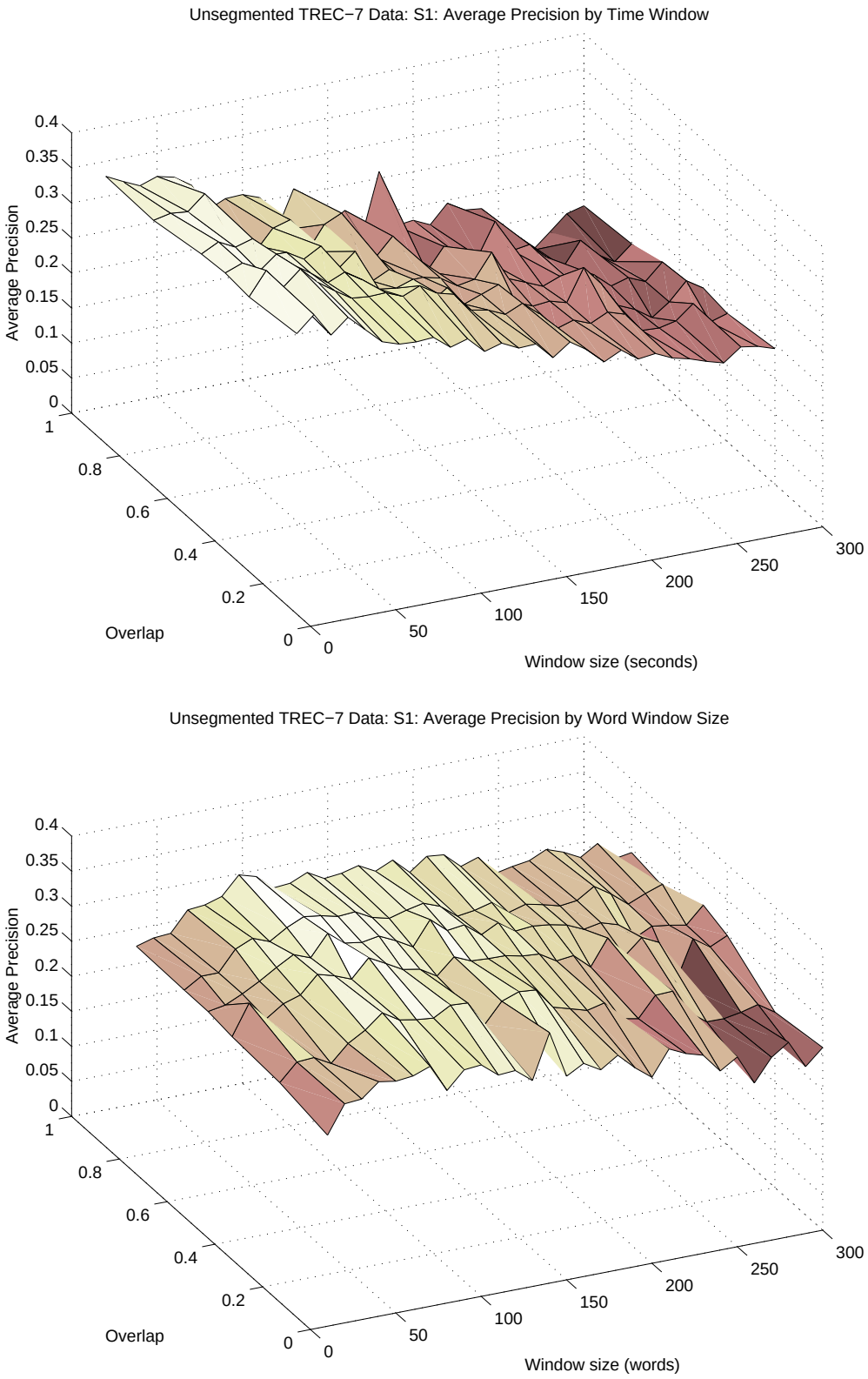

Figure 2: Effect on average precision of fixed, overlapping window automatic segmentation by time (top) and words (bottom).

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 (Allan et al. 1998). 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. As a complement to query expansion, a related approach uses similar techniques to expand the indexed documents, and has been successfully applied to spoken document retrieval (Singhal and Pereira 1999).

Robertson and Spärck Jones (1976) outlined how a query may be modified within the probabilistic model for IR that assumes term independence, given the presence of information about the relevance of documents to the query. This process is referred to as *relevance feedback*. In the purely automatic case, in which no relevance judgments are available, it is possible to approximate relevance feedback by assuming that the top R documents are relevant to the query (Croft and Harper 1979). This approach is termed blind, or pseudo, relevance feedback. Using this model, a term e from outside the query that appears in the top R documents may be ranked using a query expansion weight—sometimes called the *offer weight* (Robertson 1990)—$QEW_{RSJ}(e)$:

$$RW(e) = \log \frac{(n_R(e)+0.5)((N-n(e))-(R-n_R(e))+0.5)}{(n(e)-n_R(e)+0.5)(R-n_R(e)+0.5)} \tag{7}$$

$$QEW_{RSJ}(e) = n_R(e)RW(e)\ . \tag{8}$$

$RW(e)$ is referred to as the *relevance weight* of term e, $n_R(e)$ is the number of documents in the relevance set of R documents containing term e[10]and $n(e)$ is the number of documents in the whole collection containing e. The relevance weight is obtained from the same binomial independence model from which the collection frequency weight (1) was derived, by considering the ratio of relevant and non-relevant documents containing a term.

As well as using (8) to rank potential query expansion terms, (7) may be used to replace the collection frequency weight $CFW(e)$ in (2) or (3).[11] This may not be robust in the case of blind feedback: alternatively no additional weighting need be applied, or a weighting related to the rank of expansion terms (eg, $1/rank$) may be used.

An alternative approach to query expansion is based on term co-occurrence. Although this is outside the realm of the term independence model, QE algorithms using some notion of term correlation have been proposed. We have investigated a simplified version of the *local context analysis* (LCA) algorithm introduced by Xu and Croft (1996), referred to as LCA*. The query expansion weight $QEW_{LCA^*}(Q,e)$ for a potential expansion term e and a query Q, across a set of R (pseudo) relevant documents is defined as:

$$QEW_{LCA^*}(Q,e) = CFW(e)\sum_{t\in Q} CFW(t)\sum_{i=1}^{R} TF(e,d_i)\cdot TF(t,d_i). \tag{9}$$

[10]The $n_R(e)+0.5$ terms may be justified as a Jeffreys-Perks interpolation between the maximum likelihood estimate and the uniform (Laplace) prior, or as resulting from an expected likelihood estimate.

[11]If relevance information is not used (ie, $R = n_R(e) = 0$), and $N >> n(e)$, then $RW(e) \sim CFW(e)$.

This approach is based on the product of term frequencies weighted by the product of collection frequency weights. The original LCA formulation of Xu and Croft (1996) differed in that the term frequency correlation and the CFW of the proposed expansion term e are logarithmically compressed compared with the CFW of the original query terms t:

$$QEW_{LCA}(Q,e) = \sum_{t \in Q} \log\left(\frac{\log(\sum_{i=1}^{R} TF(e,d_i) \cdot TF(t,d_i)) \cdot CFW(e)}{\log(nr)} + \delta\right) \cdot CFW(t). \tag{10}$$

δ is a small offset (typically 0.1) introduced by Xu and Croft to avoid a zero argument to the log function.

Local context analysis does not consider distractor (non-relevant, but retrieved) documents. A discriminative term may be included by computing a similar LCA* weight over a set of distractor documents, combining with (9) using a method such as the Rocchio formula (reviewed by Harman (1992)). Experiments have indicated that adding such a discriminative term has a negligible effect. The LCA* query expansion weight (9) is used for ranking potential expansion terms only. Additional weighting can take the form of scaling the combined weights of expansion terms by $1/rank$.

8 Query Expansion Experiments

Query expansion experiments were performed using a secondary corpus of newspaper text (Washington Post and Los Angeles Times) contemporaneous with the Broadcast News data. This collection consisted of 22 471 manually segmented stories, a total of about 15.0 million words. After some development work, all experiments added a maximum of 15 expansion terms.

An initial experiment compared the effect of automatic and manual segmentation of the secondary corpus used for query expansion. The automatic segmentation used a window of 80 words, with a 50% overlap. In the case of the manually segmented secondary corpus, we assumed the top 10 documents were relevant (ie, $R = 10$); for the automatic segmentation we assumed $R = 50$. In both cases we also applied the constraint that assumed relevant documents must have a combined weight of at least $0.75\times$ that of the top ranked document. This experiment used the reference transcriptions of the broadcast news archive (R1), using the combined weight given by (3) for the manually segmented and (2) for the automatically segmented case with fixed document length. The basic RSJ algorithm (8) and the LCA* algorithm (9) were used to rank expansion terms, which were weighted by $1/rank$ in both cases.

Table 3 shows the results of the experiment comparing manual and automatic segmentation. These results indicate that there is no difference between manually and automatically segmenting the secondary corpus when using the RSJ algorithm, and that manual segmentation is slightly better when using the LCA* algorithm. Although the automatic segmentation results in a better precision at 5 documents, the precisions at 10 and 30 documents are better for manual segmentation, when using LCA* expansion, and not worse when using RSJ expansion. Since the manually segmented newswire corpus results in about 15 times fewer documents, that approach is preferred, and used for the remaining experiments in this section. The results in table 3 are the

R1: Query Expansion Experiments				
	Secondary corpus segmentation			
	Manual		Automatic	
	RSJ	LCA*	RSJ	LCA*
AveP	0.510	0.530	0.513	0.514
P5	0.565	0.557	0.583	0.574
P10	0.483	0.513	0.487	0.483
P30	0.320	0.332	0.310	0.310

Table 3: Spoken document retrieval experiments on reference transcripts (R1): Effect of manual and automatic segmentation of the secondary corpus (newspaper text). Query expansion terms were weighted by $1/rank$, a maximum of 15 expansion terms were added. For the manually segmented secondary corpus it was assumed the top 10 documents were relevant; in the case of automatic segmentation (80 words with a 50% overlap) it was assumed the top 50 documents were relevant. IR is evaluated in terms of Average Precision (AveP) and precision at 5, 10 and 30 documents (P5, P10, P30).

opposite of those we reported earlier (Abberley et al. 1999), in which automatically segmenting the secondary corpus resulted in a higher average precision. The principal difference in the earlier work was the use of the original LCA formulation (10), with log compression, to calculate the QE weight.

The next set of experiments compared the different query expansion algorithms discussed in this section using the manually segmented reference (R1) transcriptions, and both manually segmented and automatically segmented (30s window, 40% overlap) speech recognition transcriptions (S1). Again, combined weight (3) was used for the manually segmented case, and (2) was used with automatic segmentation. Following the results of the previous experiment, we used the manually segmented secondary corpus. Three query expansion algorithms were used: LCA*, RSJ and a Merge algorithm that used combined the RSJ and LCA* expansions, by adding the term weights of the two possible query expansions. Three ways of weighting expansion terms were investigated: uniform weighting, weighting based on $1/rank$ when the expansion terms were ordered by query expansion weight and (for the RSJ algorithm) the replacement of $CFW(e)$ by $RW(e)$ (computed by equation (7)) in the computation of the combined weight by (2) or (3). The results of these experiments are shown in table 4.

The results indicate that for all three cases a query expansion algorithm using $1/rank$ weighting of expansion terms results in a substantial increase in average precision over the case when query expansion was not applied. For the reference transcriptions, the improvement due to query expansion was over 10% relative; for the speech recognizer outputs, query expansion resulted in up to 20% relative improvement in the average precision. For this experiment there was a consistent ordering on the QE algorithms: Merge $>$ LCA* $>$ RSJ.

Figure 3 shows the recall-precision curves for LCA* query expansion, applied to the reference (R1) and and manually segmented speech recognizer (S1) transcriptions. It can be seen that, for recall levels below 0.8, the positive effect of query expansion more than outweighs the negative effect of a WER of 35%. Figure 4 illustrates the

R1: Query Expansion Experiments								
	No QE	RSJ			LCA*		Merge	
		Uniform	1/Rank	RW	No Wt	1/Rank	No Wt	1/Rank
AveP	0.465	0.441	0.510	0.466	0.428	0.529	0.440	0.548
P5	0.557	0.522	0.565	0.530	0.444	0.557	0.478	0.617
P10	0.435	0.413	0.483	0.457	0.422	0.513	0.409	0.526
P30	0.283	0.286	0.319	0.300	0.287	0.332	0.287	0.336

S1 Manual Segmentation: Query Expansion Experiments								
	No QE	RSJ			LCA*		Merge	
		Uniform	1/Rank	RW	No Wt	1/Rank	No Wt	1/Rank
AveP	0.407	0.410	0.487	0.417	0.395	0.506	0.395	0.509
P5	0.513	0.435	0.530	0.496	0.461	0.548	0.461	0.548
P10	0.409	0.435	0.474	0.409	0.426	0.526	0.413	0.530
P30	0.251	0.268	0.299	0.286	0.278	0.313	0.273	0.320

S1 Automatic Segmentation: Query Expansion Experiments								
	No QE	RSJ			LCA*		Merge	
		Uniform	1/Rank	RW	No Wt	1/Rank	No Wt	1/Rank
AveP	0.349	0.340	0.405	0.375	0.348	0.449	0.347	0.455
P5	0.409	0.435	0.487	0.417	0.478	0.557	0.444	0.522
P10	0.361	0.383	0.409	0.370	0.374	0.452	0.374	0.452
P30	0.232	0.235	0.264	0.252	0.242	0.299	0.241	0.290

Table 4: Query expansion experiments using R1 reference transcripts (top) and S1 speech recognizer output, manually segmented (centre) and automatically segmented using a 30s window with 40% overlap (bottom). Query expansion terms may be unweighted (Uniform), weighted according to 1/Rank or—for RSJ query expansion—the relevance weight (7) is used to replace CFW when computing the combined weight (RW).

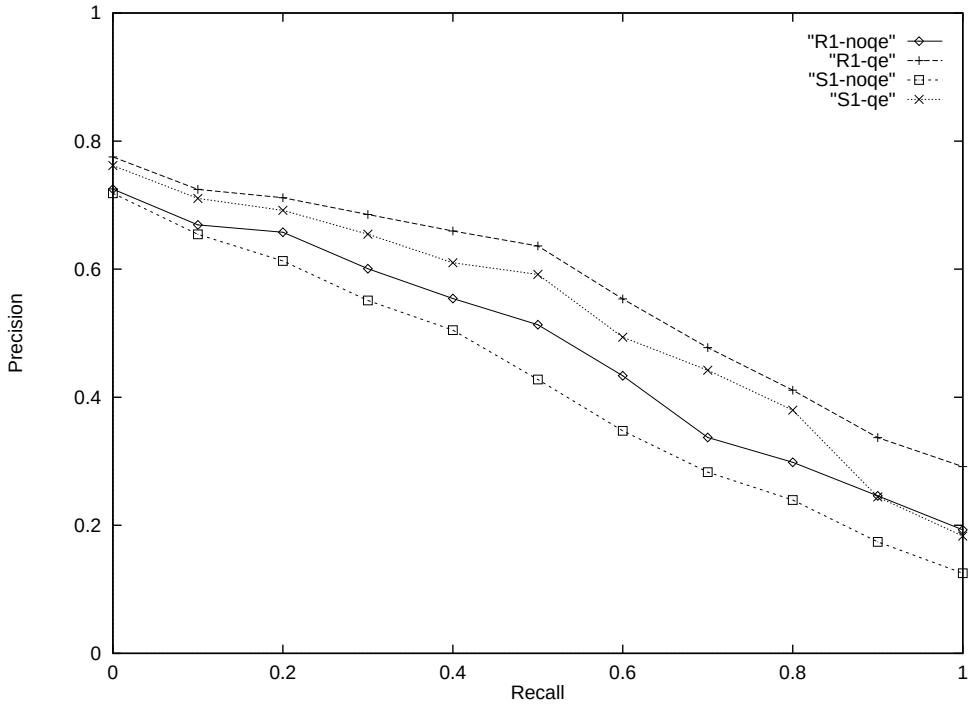

Figure 3: Effect of query expansion on recall-precision using reference transcriptions (R1) and manually segmented speech recognizer output (S1).

query-by-query change in average precision due to query expansion for the 23 queries. It can be seen that query expansion only had a significant adverse effect for 2 queries, and had a positive effect (average precision increase of over 0.1) for 9 of the 23 queries.

9 Out of Vocabulary Words

A variety of methods have been proposed to deal with the problem of OOV query words. These include the invocation of a word-spotter for OOV query words (Jones et al. 1996; Abberley et al. 1998), performing phone-level and word-level recognition and indexing on both words and phone n-grams (Witbrock and Hauptmann 1997) and dynamically choosing the lexicon according to the topic of the spoken document being recognized (Kemp and Waibel 1998). Additionally, IR techniques such as query expansion and document expansion (Singhal and Pereira 1999) can also add robustness against OOV words. In the experiments reported in this paper, we have relied on query expansion to diminish the effects of any OOV words. It turned out that only one of the words in the 23 queries was OOV with respect to our recognizer.

A set of experiments were conducted to simulate the effect of out-of-vocabulary words. These were performed by removing query words (and expanded query words) with a CFW above a threshold (either $\log(128)$ or $\log(256)$), or by removing a single query term with the highest CFW. Since the secondary corpus uses newswire text, which is not affected by OOV, these pseudo-OOV words could be used in query expansion. The results are shown in table 5. This experiment is not equivalent to choosing a

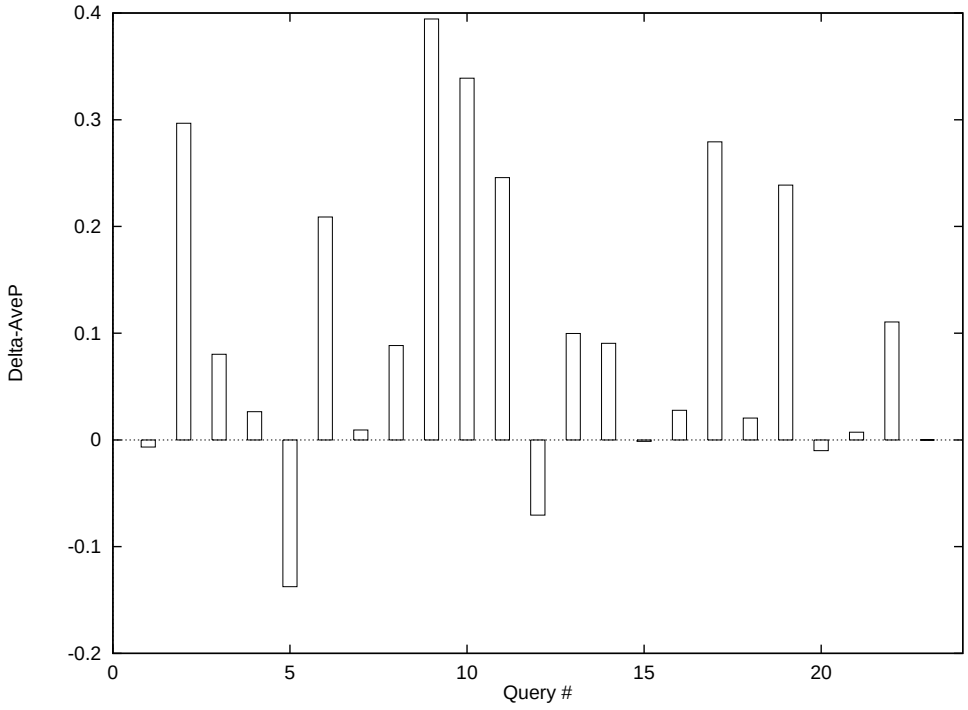

Figure 4: Query-by-query effect of query expansion in terms of change in average precision compared with no query expansion, for the manually segmented S1 case.

smaller vocabulary *a priori*, since the distribution of words in the recognized archive is used to simulate an increased OOV rate. The full language model was used at recognition time, so there was no infilling of similar sounding words. With the full 65 532 word dictionary a total of 24 562 different words are indexed; when a CFW threshold of $\log(256)$ is applied, only 9 430 words are indexed; with a CFW threshold of $\log(128)$ is applied, only 6 782 words are indexed. When the adaptive CFW threshold, based on the query, is applied the number of words indexed varies from 2 056 to 24 562 (in the case of a genuine OOV), with a median of 6 343 words. These results indicate that although simulated OOV has a deleterious effect, an acceptable average precision can still be obtained.

10 Application to BBC News

The main English language archive was constructed from six main BBC News broadcasts each day: three each from television and radio channels. This averaged about 2.5 hours of audio per day and, although by no means the full output from a newsroom, usually covered all the major breaking stories. Over the period of 1–2 years, this will result in a database that is large enough to assess the effectiveness of speech retrieval but not too onerous to manage. The system used the British English broadcast speech recognition system (as described in sections 2 and 3), and a time-based segmentation using a 30s window with 40% overlap.

At the time of writing the database contains over 800 hours of news recordings

	Average Precision	
	No QE	QE
No OOV	0.407	0.509
OOV-256	0.385	0.489
OOV-128	0.384	0.479
OOV-CFW	0.343	0.466
OOV-CFW*	–	0.419

Table 5: IR experiments simulating the effect of OOV words. OOV-256 corresponds to all words with $CFW > \log(256)$ being assumed OOV (removed from the query); OOV-128 applies a lower threshold of $\log(128)$. OOV-CFW removes the word with the highest CFW from the query; OOV-CFW* is similar to OOV-CFW, except all expansion terms with a CFW greater than that of the term removed from the query are also assumed OOV.

and it has been possible to demonstrate the system under realistic conditions. The response to these demonstrations has been encouraging and as a result, it is planned to provide more widespread access for programme researchers in the BBC archives areas. Additionally we are intending to double the number of hours of broadcast programmes recorded to include a range of others which are of interest to them.

To offer easy access and hence assess the usefulness of the system in different situations, we have created a WWW-based front-end for the user to submit searches and review results. This web-based approach will permit a wide range of users to evaluate the system using a centrally located server, dedicated to the recording and indexing tasks. In this way, it is hoped that a variety of users will be able to access the system and evaluate its effectiveness for their situation.

It is hoped that it will also be possible to run an evaluation of this system alongside the existing indexing systems that are in use in the BBC. This should give a realistic indication of the effectiveness of this approach in practical situations.

An experiment was conducted by interrogating the BBC News system with the 23 queries from TREC-7 SDR evaluation. 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 N results show that approximately half of the first five and four out of the first ten documents returned were relevant (see table 6). The corresponding figures for the TREC-7 S1 and R1 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.

11 Summary and Conclusions

We have described the development of a spoken document retrieval system, based on a real-time large vocabulary speech recognizer. We have evaluated the system on a 100 hour archive of North American broadcast news, and have applied the techniques we developed to a larger archive (currently over 800 hours) of British English broadcast

	Precision		
	BBC News	TREC-7 S1	TREC-7 R1
P5	0.505	0.513	0.557
P10	0.414	0.409	0.435

Table 6: Retrieval of BBC News application, using TREC-7 SDR queries, measured using precision at 5 and 10 documents. Query expansion was not used in these experiments.

news.

Our results have indicated that a broadcast speech recognizer with a WER of over 30% is adequate for spoken document retrieval tasks using archives of this size. Indeed, our experiments have indicated that the effect of speech recognition transcription errors can be offset (to an extent) through the use of more sophisticated IR algorithms. The WER for broadcast news is not uniform and a WER vs. time plot indicated that there is often a very large variation of WER within a story, with the planned introductory speech of the newscaster having a relatively low WER. We have also explored the issue of a finite lexicon: our speech recognition for English systems typically have an OOV rate of 1–3%, and experiment has shown that this does not impact severely on the IR performance — indeed, simulations of much higher OOV rates were also shown to have only a small effect of the average precision.

The two main IR issues that we have explored are those of automatic segmentation and query expansion. Automatic segmentation is extremely important for broadcast speech retrieval, since broadcast speech, recorded off air, does not come with document markup or story boundaries. We used simple approaches, based on overlapping fixed length windows. This approach clearly has no notion of semantics. However, with short window lengths (considerably less than a typical story length), it is possible to merge adjacent segments with large combined weights, enabling the dynamic construction of longer segments that are relevant to a given query.

We investigated two approaches for statistical query expansion both of which could be used with a secondary text corpus free from recognition errors. There was no significant difference between the two approaches; however, there was a 20% relative improvement in average precision compared with the case when query expansion was not applied.

The major caveat to the results reported here is the archive size. While an archive of 100 hours is a substantial amount of data to automatically transcribe, it results in a small archive for IR (less than one million words). Until pooled relevance evaluations can be performed on archives of at least several hundred hours, and preferably much larger, it will be difficult to arrive at firm conclusions regarding the effect of OOV words and speech recognition errors on spoken document retrieval.

Acknowledgments

This work was supported by ESPRIT Long Term Research Project THISL (EP23495). Thanks to Gary Cook for assistance with North American English broadcast speech

recognition.

References

Abberley, D., D. Kirby, S. Renals, and T. Robinson (1999). The THISL broadcast news retrieval system. In *Proc. ESCA Workshop on Accessing Information In Spoken Audio*, pp. 19–24.

Abberley, D., S. Renals, and G. Cook (1998). Retrieval of broadcast news documents with the THISL system. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 3781–3784.

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

Callan, J. P. (1994). Passage-level evidence in document retrieval. In *Proc. ACM SIGIR*, pp. 302–309.

Cook, G., K. Al-Ghoneim, D. Ellis, E. Fosler-Lussier, Y. Gotoh, B. Kingsbury, N. Morgan, S. Renals, T. Robinson, and G. Williams (1999). The SPRACH system for the transcription of broadcast news. In *Proc. DARPA Broadcast News Workshop*, pp. 161–166.

Croft, W. B. and D. Harper (1979). Using probabilistic models of document retrieval without relevance information. *Journal of Documentation 35*, 285–295.

Dharanipragada, S. and S. Roukos (1998). A fast vocabulary independent algorithm for spotting words in speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 233–236.

Ferrieux, A. and S. Peillon (1999). Phoneme level indexing for fast and vocabulary-independent voice/voice retrieval. In *ESCA ETRW on Accessing Information in Spoken Audio*, pp. 60–63.

Foote, J. T., S. J. Young, G. J. F. Jones, and K. Spärck Jones (1997). Unconstrained keyword spotting using phone lattices with application to spoken document retrieval. *Computer Speech and Language 11*, 207–224.

Garofolo, J., E. M. Voorhees, C. G. P. Auzanne, and V. M. Stanford (1999). Spoken document retrieval: 1998 evaluation and investigation of new metrics. In *ESCA ETRW on Accessing Information in Spoken Audio*, pp. 1–7.

Harman, D. K. (1992). Relevance feedback and other query modification techniques. In W. B. Frakes and R. Baeza-Yates (Eds.), *Information Retrieval: Data Structures and Algorithms*, Chapter 11, pp. 241–263. Prentice Hall.

Harman, D. K. (1996). Evaluation techniques and measures. In *Proc. Fourth Text Retrieval Conference (TREC-4)*, pp. A6–A14.

Hauptmann, A. G. and M. J. Witbrock (1997). Informedia news on demand: Information acquisition and retrieval. In M. T. Maybury (Ed.), *Intelligent Multimedia Information Retrieval*, pp. 213–239. AAAI Press/MIT Press.

Hearst, M. A. (1997). TextTiling: Segmenting text into multi-paragraph sub-topic passages. *Computational Linguistics 23*, 33–64.

James, D. A. and S. J. Young (1994). A fast lattice-based approach to vocabulary independent wordspotting. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 377–380.

Johnson, S. E., P. Jourlin, G. L. Moore, K. Spärck Jones, and P. C. Woodland (1999). The Cambridge University Spoken Document Retrieval System. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 49–52.

Jones, G. J. F., J. T. Foote, K. Spärck Jones, and S. J. Young (1996). Robust talker-independent audio document retrieval. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 311–314.

Kaszkiel, M. and J. Zobel (1997). Passage retrieval revisted. In *Proc. ACM SIGIR*, pp. 178–185.

Kemp, T. and A. Waibel (1998). Reducing the OOV rate in broadcast news speech recognition. In *Proceedings of the International Conference on Spoken Language Processing*, pp. 1839–1842.

Kraaij, W., J. van Gent, R. Ekkelenkamp, and D. van Leeuwen (1998). Phoneme-based spoken document retrieval. In *Proc. 14th Twente Workshop on Language Technology*, pp. 141–152. Universiteit Twente, Enschede, NL.

Morrison, P. and P. Morrison (1998, July). The sum of human knowledge? *Scientific American*.

Ng, C. and J. Zobel (1998). Speech retrieval using phonemes with error correction. In *Proc. ACM SIGIR*, pp. 365–366.

Ng, K. and V. Zue (1998). Phonetic recognition for spoken document retrieval. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 325–328.

Porter, M. (1980). An algorithm for suffix stripping. *Program 14*, 130–137.

Robertson, S. E. (1990). On term selection for query expansion. *Journal of Documentation 46*, 359–364.

Robertson, S. E. and K. Spärck Jones (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science 27*, 129–146.

Robertson, S. E. and S. Walker (1994). Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proc. ACM SIGIR*, pp. 16–24.

Robertson, S. E., S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford (1995). Okapi at TREC–3. In *Proc. Third Text Retrieval Conference (TREC-3)*, pp. 109–126.

Robinson, A. J. (1994). The application of recurrent nets to phone probability estimation. *IEEE Trans. on Neural Networks 5*, 298–305.

Robinson, A. J., G. D. Cook, D. P. W. Ellis, E. Fosler-Lussier, S. J. Renals, and D. A. G. Williams (2000). Connectionist speech recognition of broadcast news. *Speech Communication*. submitted.

Robinson, T., D. Abberley, D. Kirby, and S. Renals (1999). Recognition, indexing and retrieval of British broadcast news with the THISL system. In *Proc. Eurospeech*, Budapest, pp. 1067–1070.

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

Robinson, T., J. Christie, and G. Cook (2000). Time-first search for speech recognition. *Speech Communication*. submitted.

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

Siegler, M. and M. J. Witbrock (1999). Improving the suitability of imperfect transcriptions for information retrieval from spoken documents. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 505–508.

Singhal, A., J. Choi, D. Hindle, D. D. Lewis, and F. Pereira (1999). AT&T at TREC–7. In *Proc. Seventh Text Retrieval Conference (TREC-7)*, pp. 239–252.

Singhal, A. and F. Pereira (1999). Document expansion for speech retrieval. In *Proc. ACM SIGIR*, pp. 34–41.

Smeaton, A. F., M. Morony, G. Quinn, and R. Scaife (1998). Taiscéalaí: Information retrieval from an archive of spoken radio news. In C. Nikolaou and C. Stephanidis (Eds.), *Proc. Second European Digital Libraries Conference (LNCS 1513)*, pp. 429–442. Springer.

Spärck Jones, K., S. Walker, and S. E. Robertson (1998). A probabilistic model of information retrieval: development and status. Technical Report TR446, Cambridge University Computer Laboratory. http://www.ftp.cl.cam.ac.uk/ftp/papers/reports/index.html#TR446.

van Rijsbergen, C. J. (1979). *Information Retrieval* (2nd ed.). London: Butterworth.

Wechsler, M., E. Munteanu, and P. Schäuble (1998). New techniques for open vocabulary spoken document retrieval. In *Proc. ACM SIGIR*, pp. 20–27.

Williams, H. E. and J. Zobel (1999). Compressing integers for fast file access. *The Computer Journal 42*, 193–201.

Witbrock, M. J. and A. G. Hauptmann (1997). Using words and phonetic strings for efficient information retrieval from imperfectly transcribed spoken documents. In *Proc. ACM Digital Libraries '97*, pp. 30–35.

Xu, J. and W. B. Croft (1996). Query expansion using local and global document analysis. In *Proc. ACM SIGIR*, pp. 4–11.

Yamron, J., I. Carp, L. Gillick, S. Lowe, and P. van Mulbregt (1998). A hidden Markov model approach to text segmentation and event tracking. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 333–336.

Appendix : List of Acronyms

BBC	British Broadcasting Corporation
CFW	Collection frequency weight (equation 1)
CNN	Cable News Network
CW	Combined weight (equations 2 and 3)
IDF	Inverse document frequency
IR	Information retrieval
LCA	Local context analysis query expansion (equation 10)
LCA*	Modified local context analysis query expansion (equation 9)
NDL	Normalized document length
OOV	Out of vocabulary
QE	Query expansion
QEW	Query expansion weight
RSJ	Robertson-Spärck Jones query expansion (offer) weight (equation 8)
RW	Robertson-Spärck Jones relevance weight (equation 7)
R1	Reference transcripts of Hub-4 North American broadcast news
S1	Speech Recognition Transcripts of Hub-4 North American broadcast news (35% WER)
SDR	Spoken document retrieval
TF	Term frequency
TREC	Text Retrieval Conference
WER	Word error rate