Mixing and Merging for Spoken Document Retrieval

Mark Sanderson[1] and Fabio Crestani[2]

[1] CIIR, Computer Science Department
University of Massachusetts
Amherst, MA, 01007 USA
sanderso@cs.umass.edu
[2] Department of Computing Science
University of Glasgow
Glasgow G12 8QQ, Scotland
fabio@dcs.gla.ac.uk

Abstract. This paper describes a number of experiments that explored the issues surrounding the retrieval of spoken documents. Two such issues were examined. First, attempting to find the best use of speech recogniser output to produce the highest retrieval effectiveness. Second, investigating the potential problems of retrieving from a so-called "mixed collection", i.e. one that contains documents from both a speech recognition system (producing many errors) and from hand transcription (producing presumably near perfect documents). The result of the first part of the work found that merging the transcripts of multiple recognisers showed most promise. The investigation in the second part showed how the term weighting scheme used in a retrieval system was important in determining whether the system was affected detrimentally when retrieving from a mixed collection.

1 Introduction

Over the past few years the field of Information Retrieval (IR) has directed increasing interest towards the retrieval of spoken documents. Much, if not all, of the work published so far has concentrated on the use of Speech Recognition (SR) systems that identify either sub-word units (i.e. syllables [14] or phonemes [3]) or words from a limited vocabulary (i.e. the work of Jones et al [6]). Little has been published on the use of large vocabulary continuous SR systems. However, in recent years this type of system has become sufficiently accurate with a large enough vocabulary that its application to IR is feasible.

In 1997 IR and SR received a boost when the 6th Text Retrieval Conference (TREC-6) ran a Spoken Document Retrieval (SDR) track. As part of this track, a test collection of spoken documents, the SDR collection, was created: providing a common test bed for IR and SR researchers. At TREC-6 a number of presentations of work on the SDR collection were made [5]. This paper contains the work presented by the Glasgow IR group (briefly described in [2]) in addition to

other work completed more recently. It starts with an introduction to the SDR collection and the evaluation schemes used to measure retrieval effectiveness on it. This is followed by a description of the experiments conducted for TREC and the results gained from them. Next the paper describes an investigation into the issue of retrieval from a mixed collection before, finally, concluding.

2 The SDR collection

The SDR collection was created for the TREC-6 SDR track. It is composed of stories taken from the Linguistic Data Consortium (LDC) 1996 Broadcast News corpora. The collection consists of 1451 stories representing about 50 hours of recorded material. A story (document) is generally defined as a segment of news material with the same content or theme. Segmentation of the stories is performed by hand. Notice that a story is likely to involve more than one speaker and contain background music, noise, etc. The collection also comes with 49 *known item queries* (known as topics in TREC), i.e. queries for which there is only one relevant document in the collection.
The collection is supplied in a number of forms:

- digitised recordings of the broadcasts;
- detailed hand-generated transcriptions used in speech recogniser training and for speech recognition scoring containing such things as detailed timing information for the occurrence of each word. (This form of the collection should not be regarded as perfect, as there are a number of errors such as spelling mistakes.);
- the detailed transcripts with most of the recogniser training data removed leaving just a text document. Retrieval on this version of the collection provided a standard against which retrieval on the recognised collections was compared.
- automatically generated transcripts produced by a recogniser when applied to the digitised recording. A standard transcript, generated by a large vocabulary SR system from NIST/IBM, was provided with the collection to allow researchers who do not have their own recogniser a chance to experiment on such a collection. In addition to this transcript, Glasgow was given access to one produced by the Speech Group at the University of Sheffield using their Abbot large vocabulary (about 60,000 words) continuous SR system [8].

The evaluation schemes used with the SDR collection are:

- *mean rank*, i.e. the rank at which the known item was found, averaged over the queries. The smaller the number, the more effective the run;
- *mean reciprocal*, i.e. the reciprocal of the rank at which the known item was found averaged over the queries. A larger value implies better effectiveness. The mean has the range $[0, 1]$.
- *number of queries where the relevant document is found in the top n rank*, where n is 1, 5, or 10.

These measures (inevitably) have advantages and disadvantages. The mean rank may appear to be a fair measure, but a few bad retrievals easily skew the mean. For example if 48 of the 49 queries' known items were retrieved at rank position 1 but the 49th query was retrieved at rank 700, the mean rank would be 15. The mean reciprocal is not affected by this problem, but it also has drawbacks in that it is sensitive to small changes in high rank position. For example, the difference between the reciprocal of rank positions 1 and 2 is 0.5 whereas the difference between positions 4 and 5 is ten times smaller. Notice that mean reciprocal is the same as average precision when there is only one relevant document per query. The number of queries within rank position (hereby indicated as *q.w.r.p.*) is probably the easiest measure to understand and although small changes in rank position might not be reflected by it, this is most likely unimportant. The latter two measures are the ones used in the experiments presented in this paper.

3 Introduction to the experiments

A number of experiments were conducted on the SDR collection, these were used to explore different aspects of the retrieval of spoken documents. The main aim of the experiments was to find the method that produced the best effectiveness for retrieving from spoken documents. This involved comparing the effectiveness of retrieval on the two recogniser transcripts available, exploring the use of recogniser word likelihood estimates, and discovering the value of retrieving from a "merged" collection: one composed of the output of more than one speech recogniser. A second aim of the experiments was to make an initial exploration of the retrieval of so-called "mixed collections": those composed of documents resulting from different sources (e.g. speech recognition, hand transcription, optical character recognition, etc). Given the SDR collection's relative "youth" and subsequent low use, a final aim of the experiments was to examine how good it was for the task it was designed for. Observations on this aspect are made through the paper.

All of these experiments were conducted on the TREC-6 SDR collection along with the Abbot generated transcript. The SIRE systems [9] was used as the experimental retrieval system. Unless otherwise stated, throughout the experiments, the system was configured to use a $tf * idf$ weighting scheme [4]: document and query words had their case normalised, stop words were removed, and stemming (using Porter) was performed. A brief explanation of the $tf * idf$ weighting scheme is reported in section 3.1.

The two recognisers vs. the hand transcription

The initial experiment was to discover which recognised transcript produced the best retrieval effectiveness and how different that effectiveness was from that obtained using the hand-transcribed version. Using SIRE in its standard set up, it was found that across all evaluation measures, the Abbot transcript

was better than NIST/IBM, in fact retrieval from the Sheffield transcript was almost as good as retrieval from the hand transcription (a strong indication of the utility of using this type of SR system). It is worth noting, however, that the SDR retrieval tasks are rather easy, as even on the poorest configuration, for 41 of the 49 known item queries, the item was retrieved in the top ten.

The differences between the two recognisers were attributed to differences in their accuracy. An analysis of the Word Error Rate (WER) was performed on the Abbot and NIST/IBM transcripts, the results of which are shown in the following table. Unlike a classic WER that is computed over all words, this rate was calculated for the SDR query words (after stop word removal) as well. As can be seen, from table 1 the Abbot transcripts were more accurate than NIST/IBM.

Table 1. Word Error Rates (%) for the two recognisers used.

	Stop Words	*Query Words*	*Other Words*
Abbot	40.3	33.1	39.7
NIST/IBM	49.4	45.5	49.0

Table 2 reports a comparison of the effectiveness of the three different SR systems using the same weighting scheme; as it can be seen Abbot performs better than NIST/IBM for lower values of the q.w.r.p.

3.1 Additional recogniser output

Unlike the transcript available from NIST/IBM, the transcript available to us from Abbot contains a value attached to each word that is an indication of Abbot's "confidence" of recognising a word. It was speculated that this additional information might be incorporated in a term weighting scheme to improve retrieval effectiveness: i.e. words that had a higher confidence value were more likely to be correctly recognised and, therefore, should be assigned a high term weight.

The value attached to a particular word in a document was regarded as a probability indicating the likelihood of that word being spoken in that document. This required the values to be mapped into the range $[0, 1]$. A number of different mappings were used, as reported in [1], and the retrieval results were compared with those obtained by discarding this additional information. This probability was incorporated into a tf weight to produce a probabilistic term (ptf) weight. Therefore, given a document d_i represented by means of a number of index terms (or words) t_j, the $tf * idf$ weighting scheme is defined as:

$$tf * idf(d_i) = \sum_{j=1}^{n} tf_{ij}(C + idf_j)$$

Table 2. Comparison of effectiveness for three different SR systems.

Hand trans. with tf_idf_porter_stop	
Mean reciprocal	0.704332
q.w.r.p. = 1	27
q.w.r.p. *leq* 5	43
q.w.r.p. *leq* 10	46
NIST/IBM rec. with tf_idf_porter_stop	
Mean reciprocal	0.610862
q.w.r.p. = 1	23
q.w.r.p. *leq* 5	38
q.w.r.p. *leq* 10	41
Abbot rec. with tf_idf_porter_stop	
Mean reciprocal	0.690164
q.w.r.p. = 1	31
q.w.r.p. *leq* 5	36
q.w.r.p. *leq* 10	41

where tf_{ij} is the frequency of term t_j in document d_i, idf_j is the inverse document frequency of term t_j, and C is a constant that is set experimentally to tailor the weighting schema to different collections. The values of tf_{ij} and idf_j are defined as follows:

$$tf_{ij} = K + (1 - K)\frac{freq_i(t_j)}{maxfreq_i}$$

where K is a constant that need to be set experimentally and $maxfreq_i$ is the maximum frequency of any term in document d_i, and

$$idf_j = \log \frac{N}{n_j}$$

where N denotes the number of documents in the collection and n_j the number of documents in which the term t_j occurs. This weighting scheme has been used extensively in the experiments reported in this paper.

The $ptf * idf$ uses the probabilities given by Abbot to evaluate $freq_i(t_j)$ as:

$$freq_i(t_j) = \sum_{d_i} Prob(t_j)$$

We used this new values of frequencies in the above tf formula to produce so called ptf values to be used in the $ptf * idf$ weighting scheme.

The experiment to examine if the ptf weight would improve effectiveness was a simple comparison between a retrieval system using a $tf * idf$ weighting scheme against the $ptf * idf$ scheme. As can be seen in the following table,

the comparison showed the ptf weighting scheme to be inferior to the simpler $tf*idf$. Although not shown in table 3, a number of transformations were used to map the likelihood values into a probability, all other mapping produced worse retrieval effectiveness than the scheme shown here.

Table 3. Results of experiments using $ptf * idf$ weighting.

Abbot rec. with ptf_idf_porter_stop	
Mean reciprocal	0.665091
q.w.r.p. = 1	29
q.w.r.p. *leq* 5	35
q.w.r.p. *leq* 10	41

An analysis was conducted to see why the additional likelihood data was detrimental to effectiveness. It was realised that the likelihood value attached to a particular word was generally higher the longer a word was. Therefore, the likelihood data should have been normalised to the length of the word, measured in letters or duration to speak it. However, a variation of this technique has already been tried by Siegler et al [10] with no success. It would appear that the reason for the lack of utility of this data remains to be discovered.

4 Experiments with merged collections

The previous section presented an investigation of the use of probabilities assigned by Abbot to words in the transcription. This work led us to consider if there was some other way of generating confidence values to assign to recognised words. The two speech transcripts (NIST/IBM and Abbot) were quite different from each other as the following example illustrates.

NIST/IBM:

```
..I will talk about blacks and winds we
eventually go wrong a of the tough
question who he hid...
```

Abbot:

```
..we talked about blanks and whites we
eventually get around to the tough
question his own unions say well....
```

Hand generated transcript:

```
..when we talk about blacks and whites we
eventually get around to the
tough question some of you are...
```

It was realised that by using a simple strategy of concatenating the documents of the two transcripts, one would effectively produce a collection with word confidence information. If, for example, the two documents fragments shown above were concatenated, the correctly recognised word "question" would occur twice, but the incorrectly recognised word "winds" would only occur once. Through use of a tf weighting scheme, "question" would receive a higher weight than the word "winds". It was decided to test if this strategy of concatenation, or merging, of the documents improved retrieval effectiveness.

Of course, such a merged collection would contain two separate hypotheses on what was spoken and would therefore, contain more correctly (and incorrectly) recognised words. Regardless of the tf producing confidence values for words, the mere presence of the extra words might, on their own, improve retrieval effectiveness and it was decided that this should also be tested.

Table 4 shows the results of retrieval on the merged collection using a $tf * idf$ weighting scheme, as can be seen when compared to the retrieval results presented in the previous tables, the merged strategy was slightly better (on three of the four measures), though very similar, to retrieval on the Abbot transcripts. From this result it would appear that merging the good transcript with the poorer has not reduced effectiveness and possibly improved it.

Table 4. Results of experiments the merged collection using $tf * idf$ weighting.

Merged with tf_idf_porter_stop	
Mean reciprocal	0.699470
q.w.r.p. = 1	30
q.w.r.p. *leq* 5	41
q.w.r.p. *leq* 10	42

Notice, the utility of merged collections was also investigated by two other groups at TREC-6: Siegler et al and Singhal et al [11]. Both were merging the NIST/IBM standard transcript with the output of their own SR systems. Both reported similar results on merging to those made here: marginal improvements in effectiveness were found.

A further experiment using the merged collection was to try to discover if any benefit from it was from the larger vocabulary within it or from the (presumably) better tf weights resulting from the combination of the two recogniser hypotheses. To discover this, SIRE was re-configured to ignore tf weights and use only idf term weighting when retrieving. As can be seen in table 5, retrieval experiments were conducted on the merged collection and its two component collections. Here, in contrast to the previous experiment, effectiveness from the merged collection was similar but slightly worse than effectiveness on the Abbot transcript. Perhaps this indicates that removal of tf weights from the merged

collection was detrimental, but, the differences are so small that nothing conclusive was drawn from this result.

Table 5. Results of experiments the merged collection using *idf* weighting.

Merged with idf_porter_stop	
Mean reciprocal	0.593621
q.w.r.p. = 1	24
q.w.r.p. *leq* 5	35
q.w.r.p. *leq* 10	37
NIST/IBM rec. with idf_porter_stop	
Mean reciprocal	0.587153
q.w.r.p. = 1	24
q.w.r.p. *leq* 5	35
q.w.r.p. *leq* 10	37
Abbot rec. with idf_porter_stop	
Mean reciprocal	0.606046
q.w.r.p. = 1	23
q.w.r.p. *leq* 5	38
q.w.r.p. *leq* 10	40

5 Experiments with mixed collections

One other area of investigation afforded by the SDR collection was an opportunity to investigate the retrieval of documents from a mixed collection: one composed of both hand-transcribed and recognised documents. Here the focus was on whether one type of document (transcribed or recognised) was more likely to be retrieved over the other and to discover if such a preference was affected by the term weighting scheme used.

5.1 Previous work

There appears to be little previous work on the topic of retrieving from mixed document collections. Researchers have, however, investigated the manner in which retrieval functions are affected by errors in recognised documents.

Concentrating on documents recognised by an OCR system, Taghva et al [13] and Singhal et al [12] both found that OCR error did not impact on effectiveness greatly, but found that existing schemes for ranking documents could be adversely affected by recognition error. Singhal et al reported that if the OCR system they were using incorrectly recognised a letter in a word (e.g. "systom"

for "system"), this would result in a word that was likely to be rare and, therefore, have a high idf. Due to the manner in which document rank scores are calculated in the vector space model, the presence of a number of such "error words" in a document would result in its being ranked lower than it if it were without error. Buckley presented an alteration to the vector space model that addressed this problem. Tahgva et al found a similar form of problem with the length normalisation part of the INQUERY term weighting scheme. If an OCR system incorrectly recognised an image as being a textual part of a document, a large number of extra 'words' were introduced and the length of said document was increased by a large amount.

This research has shown how recognition error can adversely affect the ranking of documents. Although the research presented examined OCR, one can imagine similar types of problems arising in speech recognition: incorrectly recognising a word as another; or trying to recognise a sound that is not speech. In the context of mixed document retrieval, documents containing these errors are likely to cause similar ranking problems to those reported in the research presented above. However, other forms of error may affect retrieval in the context of a mixed document collection and, therefore, an experiment was conducted to investigate this.

5.2 The experiments

To conduct the investigation the hand-transcribed and the Abbot recognised collections were combined into a single collection of 2902 documents. (The Abbot transcript was used, as it was more accurate than the NIST/IBM.) A retrieval experiment was conducted the measurement of which concentrated on where in the ranking the hand-transcribed and recognised documents were to be found. To establish this, two measures of retrieval effectiveness (using mean reciprocal) were made, one based on the location of the relevant hand-transcribed documents and one on the location of the relevant recognised documents. Any difference between these two measures was taken to indicate the different rank positions of the two document types.

The configuration of the first experiment used $tf * idf$ weighting. Results of this experiment are shown in the following table 6.

Table 6. Results of retrieval experiments using $tf * idf$ weighting.

	Hand Rel.	*Abbot Rel.*
Mean reciprocal	0.591955	0.380770

As can be seen, there is a large difference between the two figures. This result was interpreted as showing that the hand-transcribed documents were being retrieved in preference to the recognised. It was speculated that the reason for

this difference was caused by the terms in the recognised documents generally having a smaller tf than that found in the hand-transcribed documents. In other words, a query term found to occur five times in the hand-transcribed version of a document might only be correctly recognised twice in the spoken version. Therefore, the term in the recognised document would have a lower tf than in the transcribed document and this would lower the relevance score assigned to the recognised document. In order to test this speculation the first experiment was repeated using just idf weighting. The results of this are shown in table 7. As can be seen, the difference between the two figures was much smaller, indicating that it was the differences between the tf weights that caused the preference of retrieving the hand-transcribed documents.

Table 7. Results of retrieval experiments using idf weighting.

	Hand Rel.	*Abbot Rel.*
Mean reciprocal	0.479931	0.496340

Clearly these results reveal a shortcoming of the $tf * idf$ weighting scheme we have adopted within our IR system. We suspect, however, that this may be a problem for many such weighting schemes as most make the implicit assumption that term frequencies within the documents of a collection are distributed similarly across that collection. A means of handling this situation may be sought in the work of Mittendorf [7] who has examined the issue of retrieval from corrupted data.

6 Conclusions and future work

The work presented in this paper was very much an initial foray into the field of IR using large vocabulary SR. However, we feel that the experimental results presented here give an indication to some of the issues and potential solutions in this area.

First, the utility to IR of large vocabulary continuous SR systems like Abbot has been demonstrated through the retrieval results gained on the SDR collection. However, results as good as these may not be entirely fair as very few of the queries in SDR had words outside Abbot's vocabulary (whether these queries were created with the vocabularies of SR systems in mind is unknown). More "realistic" queries containing many proper nouns might produce different results and require an alternative approach: for example, an SR system using both word and sub-word unit recognition; or an IR system using a query expansion technique to expand, from a text corpus, unrecognised query words with those in the SR system's vocabulary (using, for example, Local Context Analysis [15]).

From the experimental results, it is clear that more work is required if the use of likelihood values will improve retrieval effectiveness. More promising is the

use of merged collections that showed some slight improvement in effectiveness. Finally, the experiments on mixed collections showed that care must be taken in selecting a weighting scheme that handles the different term occurrence statistics of documents taken from different sources.

References

1. F. Crestani and M. Sanderson. Retrieval of spoken documents: first experiences. Research Report TR-1997-34, Department of Computing Science, University of Glasgow, Glasgow, Scotland, UK, October 1997.
2. F. Crestani, M. Sanderson, M. Theophylactou, and M. Lalmas. Short queries, natural language and spoken document retrieval: Experiments at Glasgow University. In *Proceedings of TREC-6*, Gaithersburg, MD,USA, November 1997. In press.
3. C. Gerber. The design and application of an acoustic front-end for use in speech interfaces. M.Sc. Thesis, Department of Computing Science, University of Glasgow, Glasgow, Scotland, UK, February 1997. Available as Technical Report TR-1997-6.
4. D. Harman. Ranking algorithms. In W.B. Frakes and R. Baeza-Yates, editors, *Information Retrieval: data structures and algorithms*, chapter 14. Prentice Hall, Englewood Cliffs, New Jersey, USA, 1992.
5. D. Harman, editor. *Proceedings of the Sixth Text Retrieval Conference (TREC-6)*, Gaithersburg, MD, USA, November 1997. (In press.).
6. G.J.F. Jones, J.T. Foote, K. Spark Jones, and S.J. Young. Video mail retrieval using voice: an overview of the Stage 2 system. In *Proceedings of the MIRO Workshop*, Glasgow, Scotland, UK, September 1995.
7. E. Mittendorf and P. Schauble. Measuring the effects of data corruption on information retrieval. In *Proceedings of the SDAIR 96 Conference*, pages 179–189, Las Vegas, NV, USA, April 1996.
8. T. Robinson, M. Hochberg, and S. Renals. The use of recurrent networks in continuos speech reognition. In C.H. Lee, K.K. Paliwal, and F.K. Soong, editors, *Automatic Speech and Speaker Recognition - Advanced Topics*, chapter 10, pages 233–258. Kluwer Academic Publishers, 1996.
9. M. Sanderson. System for information retrieval experiments (SIRE). Unpublished paper, November 1996.
10. M.A. Siegler, M.J. Witbrock, S.T. Slattery, K. Seymore, R.E. Jones, and A.G. Hauptmann. Experiments in spoken document retrieval at CMU. In *Proceedings of TREC-6*, Gaithersburg, MD, USA, November 1997.
11. A. Singhal, J. Choi, D. Hindle, and F. Pereira. AT&T at TREC-6: SDR Track. In *Proceedings of TREC-6*, Washington DC, USA, November 1997.
12. A. Singhal, G. Salton, and C. Buckley. Lenght normalisation in degraded text collections. Research Report 14853-7501, Department of Computer Science, Cornell University, Ithaca, NY, USA, 1995.
13. K. Taghva, J. Borsack, and A. Condit. Results of applying probabilistic IR to OCR. In *Proceedings of ACM SIGIR*, pages 202–211, Dublin, Ireland, 1994.
14. M. Wechsler and P. Schauble. Speech retrieval based on automatic indexing. In *Proceedings of the MIRO Workshop*, Glasgow, Scotland, UK, September 1995.
15. J. Xu and W.B. Croft. Query expansion using local and global document analysis. In *Proceedings of ACM SIGIR*, pages 4–11, Zurich, Switzerland, August 1996.