A SYSTEM FOR UNRESTRICTED TOPIC RETRIEVAL FROM RADIO NEWS BROADCASTS

David A. James
UBILAB
Union Bank of Switzerland
Bahnhofstrasse 45
CH-8021 Zurich
Switzerland

ABSTRACT

The "topic classification" systems described in the speech literature typically partition a collection of spoken messages into a small number of pre-defined topics. As such, they are only useful if the set of message topics does not vary over time. However, the techniques of textual information retrieval (IR) have long allowed for retrieval by arbitrary subject from a document collection. This paper describes experiments in unrestricted retrieval from a collection of radio news broadcasts. A hybrid message indexing strategy, with conventional word recognition and a fast lattice-based wordspotter, allows for the retrieval of news reports concerning any subject. The results show that retrieval can be carried out extremely quickly and that high accuracy is possible, even with errorful recognition output.

1. THE MESSAGE RETRIEVAL PROBLEM

There is considerable interest, in the speech research community, in the automatic classification of spoken-word recordings solely by their acoustic content. *Topic classifiers* achieve this by pre-defining a group of allowable topics, and training weights which relate the recognition of certain words to topic membership [1][2]. This approach guarantees good performance at the expense of flexibility. It is no trivial matter to add new topics, or split an existing one to create a finer granularity of classification. The utility of such classifiers is therefore limited.

Recently, there have been attempts to integrate the well-understood methods of textual information retrieval (IR) [3] into an acoustic recognition front-end [4][5]. IR allows for retrieval from a text collection in response to a user's arbitrary information *request* (e.g. "get me items about the United Nations"). The main problem of applying such an approach to spoken messages is that the vocabulary of interest is not available in advance, and not predictable. Of the systems for which results have been published, the Cambridge VMR system [4] has until now been restricted to indexing messages using a fixed keyword set. Schäuble and Glavitsch [5][6], and Wechsler and Schäuble [7], have proposed novel sub-word indexing methods, but published results have until recently been confined to simulation.

This paper describes experiments in unrestricted retrieval from a collection of spoken news reports from BBC Radio 4 in the UK. Subsequent sections describe the collection of the acoustic corpora and the sequence of experiments.

2. THE SPOKEN MESSAGE COLLECTION

Several factors suggested the use of speech data obtained from BBC Radio 4 news reports [8]. The typical format of these broadcasts is as follows; the newsreader first gives the headlines, and then, for each story, gives a concise summary and introduces a more in-depth, "on the spot" report from a journalist. The regular rotation of newsreaders meant that the same small group of speakers could be used for both acoustic model training and message retrieval experiments. Also, the newsreader speech is read as opposed to spontaneous, and clearly spoken, as the newsreaders are all professional announcers.

In total, 3650 sentences of acoustic training data were collected over a three-week period in August 1993 from seven speakers, 4 male and 3 female. Each sentence was manually endpointed and orthographically labelled. A phonetic dictionary was used to convert the orthographic labels to phoneme sequences. The total size of the training data collection was 5.5 hours, with 47 minutes of speech from each speaker. The spoken message test collection was collected over a two-week period in January 1994. A total of 337 message files, each corresponding to a single news item, were collected. The test collection totalled 2 hours 27 minutes in length, with the same seven newsreaders appearing in both test collection and training corpus.

3. INITIAL EXPERIMENTATION

Information Retrieval (IR) is the name given to the set of techniques that have been developed to address the problem of locating textual items of interest in a large document collection [3]. The core methods of IR are *document indexing*, and *query-document matching*. On addition to a collection, a document is processed to yield a *document description* which stands in for it during retrieval. The process typically involves the removal of function words and word suffixes. The creation of the description is known as the document indexing process. The *request*, the user's expression of information need, is processed in the same way to produce a so-called *query*. The query is then matched against each document description to yield a query-document *score*. This score is typically related to the number of terms in common between the query and the document. The documents may then be ranked by this score, and the ability of the system to retrieve relevant documents thereby assessed. Before retrieval experiments can be performed, *relevance assessments* - subjective partitions of the collection into subsets of relevant and non-relevant documents - are made for each request. These assessments act as a fixed 'snap-shot' of information need, thereby allowing differing retrieval methods to be compared fairly. Since the methods of IR circumvent the fixed-topic restriction, they allow retrieval on a non-static collection - for example, a collection of news reports updated daily.

Forty areas of interest (‘subjects’), a number of which are illustrated in Table 1, were identified in the message collection. Requests and relevance assessments were supplied for each subject by 8 volunteers. It should be clear that the sets of relevant messages do not partition the collection exhaustively; several messages may be assessed as relevant to more than one request, and some as relevant to none. The requests were processed to generate the set of queries, which contained a total of 206 distinct terms. The average query size was 6.7 terms, and an average of 11.75 messages were judged relevant to each request.

3.1 Text-Based Retrieval

First, retrieval was performed using the message text transcriptions. Text retrieval performance has previously been useful in calibrating spoken message retrieval results [4]. The conventional IR technique of *suffix stripping* was applied to both query and message terms to reduce word form variation [9]. Instead of the simplest possible scoring method, i.e. counting the number of common terms between a query and a message, an extended scoring metric was used. Terms were *weighted* to reflect their relative frequency across the entire set of messages, and within each message. The weight w_{ij} of term i in message Mj took the form

$$w_{ij} = (0.5 + 0.5\frac{tf_{ij}}{\max\limits_{k \in M_j} tf_{kj}}).\log(\frac{N}{n_i})$$

where tf_{ij} was the frequency of term i in message M_j, n_i the number of messages containing term i, and N the total number of messages in the collection [10]. The query-message score was calculated by summing the weights for the query terms occurring in the message. As in [4] and [6], retrieval accuracy for the entire set of requests was measured in terms of the *average precision* [3]. Precision is defined as the proportion of relevant documents retrieved; the average precision is calculated first by obtaining, for each query, an average of the precisions measured after each relevant document is retrieved, and then by averaging this figure across all queries. The average precision of retrieval on the textual transcriptions of the message collection was found to be 0.7038; that this is quite a high figure, compared to results on text collections, is due mainly to the relatively small number and concise nature of the messages. In comparison, experimental text collections typically contain millions of documents.

Subject	Query
Train Crashes	*accidents crashes derailed railways trains*
South Africa	*africa elections mandela reform south*
Earthquakes	*disasters earthquakes faultlines richter*

Table 1: Three subjects with their corresponding queries.

3.2 Speech-Based Retrieval

The first *spoken* message retrieval experiment involved the use of a Viterbi wordspotter to detect the 206 query terms in the message collection. The acoustic training data was parametrised, using the HTK toolkit [11], to generate 12 mean-compensated MFCCs, plus 1st and 2nd differential coefficients and log energy, every 10ms. The HTK training tools were then used to estimate 47 12-mixture Gaussian monophone HMMs. The message collection was parametrised identically. It was not possible to train whole-word models for any of the query terms, since the occurrence of sufficient training exemplars could obviously not be guaranteed. Instead, term models were built by concatenating monophones, and placed in a parallel network with a garbage model. The garbage model was built from the monophone states using an agglomerative clustering procedure which allows the garbage model accuracy to be precisely controlled [12]. In this case, the number of garbage model states was set to allow a large number of term hypotheses per message, and the term log likelihood scores post-processed to obtain the durationally-normalised log likelihood ratio score (DNLLR) [13]. Setting a single, keyword-independent threshold on this score allowed the term correctness/false alarm trade-off to be varied, and retrieval experiments to be performed at differing trade-offs.

The best average precision obtained here was 0.5304, or, when expressed as a ratio of the corresponding figure for text, 75.36%. This was obtained at a DNLLR threshold of -1.5. At higher values of this threshold, term detections were more certain but fewer in number; at lower values, a greater number of correct detections was recorded but the expense of many ‘contaminating’ false alarms. The optimal output, as far as retrieval is concerned, unsurprisingly occurs somewhere between the two extremes [4].

So far, messages are indexed solely at ‘query-time’ - that is, the moment the queries are submitted to the retrieval system, and therefore the moment that the query terms are first known. The approach is rather unsatisfactory in a number of areas. Firstly, it was necessary to ‘batch’ queries, that is, detect all 206 terms from the 40 queries in a single pass; this would almost certainly not be acceptable to users of a practical system. Also, this pass took 16 hours of CPU time on a SGI workstation. Moreover, optimal retrieval performance depends on the precise setting of the DNLLR threshold, which itself depends on unpredictable factors, such as the extent to which each term indexes relevant messages, and ease of term detection.

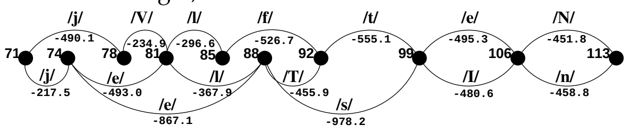

Figure 1: A phone lattice of degree 2 for the word *yeltsin.*

The next experiment was based on a two-stage approach to message indexing. In the first stage, phone recognition using a modified Viterbi recogniser generates, for each message, a phone *lattice* encoding a number of paths, each labelled with a phone hypothesis and acoustic log likelihood score, at every point. Figure 1 illustrates an example of a lattice. At query-time, each term is detected by searching each precomputed lattice for the exact phone sequence corresponding to the pronunciation of that term [12]. The advantage of the method is that messages can be indexed at query-time far more quickly by lattice wordspotting than by Viterbi wordspotting. In addition, since each message lattice contains, by definition, the Viterbi phone sequence, term detections can be assigned the DNLLR score with no extra computation. Pronunciations for some of the query terms were in this case supplied manually; however, recent experiments have shown the effectiveness of letter-to-sound rules to generate phone sequences for unknown words [14].

In this experiment, query term occurrences were detected in message lattices of degree 8 (i.e. with at most 8 phone edges converging and diverging at every speech frame). Messages were then retrieved at varying DNLLR thresholds. Table 2 summarises the results and compares them with the earlier method. The best average precision, 0.4971 (70.63% of text) was found to occur at the same threshold as earlier. Although precision is lower than in the previous experiment, the only query-time delay here was due to lattice term detection. With the set of lattices pre-loaded into memory, this delay averaged around 8 CPU seconds per query, or 0.5 CPU seconds/term/hour of speech, on the SGI workstation. Since the new method involves no query-dependent acoustic wordspotting, it is a far more realistic approach to the message retrieval problem than the earlier Viterbi method.

The figures of merit (FOMs) for the Viterbi and lattice wordspotters on the 206-term keyword set were 67.79% and 60.46% respectively; however, the FOM is not such a useful statistic here, since it does not reflect the fact that a large number of false alarms of a spurious query term, which does not appear in the message collection at all, may significantly decrease retrieval precision.

	Viterbi WS	Lattice WS
Recognition network	206 terms	47 phones
Wordspotter FOM	67.79%	60.46%
Average Precision	0.5304	0.4971
% of text Avg. Prec.	75.36%	70.63%
Query-Time Search	24 mins (batch mode)	8 secs

Table 2: Initial retrieval results

4. MODELLING IMPROVEMENTS

Although the lattice method offers a workable approach to spoken message retrieval, it has a number of deficiencies. Since there is no explicit garbage model, the detection of short query terms is quite inaccurate. Also, no attempt is made to exploit the fact that, as with many IR text collections, the messages belong to a single domain of discourse. The adoption of a *hybrid* indexing strategy addresses both these issues. This approach augments the lattice method with an initial, conventional word recognition pass using a domain-specific language model. While it cannot model every word appearing in the collection, the language model should allow for the detection of a significant number of terms likely to occur in the radio news domain, and therefore a significant number of query terms. This information is then *cached*, so that at query-time, only uncached query terms, those not appearing in the language model, have to be detected in the lattices. This approach will not only improve the detection of short terms, which are likely to appear in the language model, but also cut the query-time lattice search time.

The hybrid strategy was tested with a 3000-word *bigram* language model, estimated from a frequency ranking of words in a corpus of news stories from a UK newspaper [15]. The word-pair probabilities were smoothed using Katz' "back-off" method [16]. Suffix-stripping was applied to query terms and word recogniser output alike.

It was unnecessary to build an acoustic model for out-of-vocabulary (OOV) words. Since the bigram applies a local constraint to the word sequence, rather than a global one, the incorrect recognition of an OOV word as a known, language-model word is unlikely to impact strongly on the recognition of the message as a whole. Moreover, as Kupiec et al. [17] pointed out, as long as the queries contain enough terms, query-message matching acts as a semantic filter on the recognised messages; it picks out those messages in which a high number of semantically-related query terms appear, thereby minimising the effect of semantically-unrelated substituted terms.

It can be seen from the middle column of Table 3 that the hybrid method offers a significant precision increase over both earlier methods, as well as decreasing the query-time search requirement. Of the 206 query terms, 152 were contained in the 3000-word vocabulary. In addition, Figure 2 shows that the performance is now more independent of the DNLLR term detection threshold. It is only necessary to apply the threshold to the lattice term detections; since the bigram sharply reduces the number of term false alarms, there are few low-scoring hypotheses of the 152 in-vocabulary terms to threshold away.

	Monophone	Tri/Biphone
Word Corr/Acc	65.40%/55.20%	70.62%/56.92%
Wordspotter FOM (on 54 terms)	62.36%	71.85%
Average Precision	0.5955 (84.61%)	0.6512 (92.52%)
Query-Time Search	3.2 secs	3.2 secs

Table 3: Results with the hybrid retrieval model.

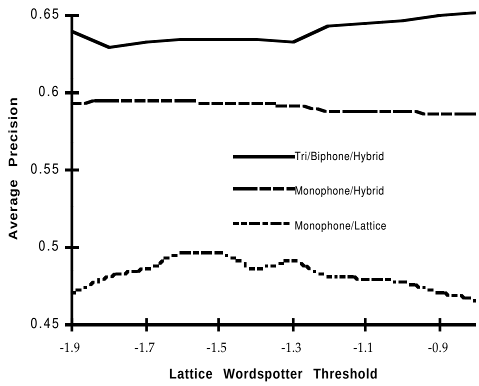

Figure 2: Precisions at varying wordspotter thresholds for the lattice-only and hybrid retrieval systems.

A second improvement to the retrieval system was made by using context-dependent, state-clustered phone models [14]. The HTK tools were used to estimate two sets of 8 Gaussian mixture models from the acoustic training data; a set of 6857 state-clustered, word-internal triphones for the word recogniser, and a set of 1936 right biphones for message lattice generation. Models for triphones and biphones unseen in the training data were synthesised using the phonetic decision trees generated during state clustering. As the third column of Table 3 shows, improving the acoustic models significantly increased word recogniser correctness and lattice wordspotter FOM. Average precision finally reached 92.52% of the textual level.

5. RELEVANCE FEEDBACK

Relevance feedback, a popular IR technique, is the training process whereby an initial query is iteratively reformulated after the assessment of retrieved documents. It allows for re-weighting of existing query terms and the incorporation of new terms into the query. A proven weighting scheme, which takes into account the occurrence of a term in relevant and non-relevant documents, is the predictive probabilistic term relevance weight of Robertson and Sparck Jones [18];

$$tr_i = \log \frac{(r_i + 0.5)(N - n_i - R + r_i + 0.5)}{(R - r_i + 0.5)(n_i - r_i + 0.5)}$$

where N is the number of assessed documents, n_i the number of them containing term i, R is the number of assessed documents that are relevant, and r_i is the number of them which contain term i. This is equivalent to the quantification of the dependency between a term and a topic which is central to topic classifiers.

An experiment was performed to determine the utility of relevance feedback in spoken message retrieval. A set of concise queries, with an average of 3.8 terms per query, was obtained from the 8 volunteers and used to retrieve messages. Two passes were performed; a reference pass using text transcriptions, and one using the triphone/biphone hybrid indexing method. For each pass and for each initial query, the above term relevance weight was calculated for each term occurring at least once in the set of the top 10 retrieved messages. The terms were then ranked by weight, and the top ten taken as the new query. Messages were then retrieved, using the new query, from the remainder of the collection. Terms were now weighted by substituting the probabilistic relevance weight tr_i into the formula for w_{ij}, in place of the earlier log term. Table 4 shows that relevance feedback can increase the average precision of retrieval for the entire message collection, and that the percentage improvement observed for spoken messages is only slightly smaller than for textual messages.

	No Feedback	With Feedback	Increase
Text	0.5444	0.5886	8%
Speech	0.4737	0.5100	7%

Table 4: Results with the use of relevance feedback.

6. CONCLUSION

The well-understood techniques of textual IR can be applied to the problem of spoken message retrieval. The use of the hybrid indexing method offers total search term flexibility without sacrificing speed, and retrieval can be performed with over 90% of the accuracy of a textual reference run. In addition, the IR method of relevance feedback can automatically generate an expression of information need from an arbitrary initial collection of search terms. Further work will examine the potential of sub-word units for indexing larger message collections.

ACKNOWLEDGEMENTS

The work described in this paper was performed at Cambridge University Engineering Department under the supervision of Prof. S. J. Young and with funding from the Engineering and Physical Sciences Research Council and Downing College. The author would like to thank Drs. K. Sparck Jones, J. T. Foote and G.J.F. Jones for their help.

REFERENCES

[1] R. C. Rose, E. I. Chang and R. P. Lippmann. Techniques for Information Retrieval from Voice Messages. In Proc. ICASSP, pp317-320. IEEE, Toronto, 1991.
[2] J. McDonough, K. Ng, P. Jeanrenaud, H. Gish and J. R. Rohlicek. Approaches to Topic Identification on the Switchboard Corpus. ICASSP, pp I-385-388. IEEE, 1994.
[3] C. J. van Rijsbergen. Information Retrieval (2nd edition). Butterworths, London, 1979.
[4] G. J. F. Jones, J. T. Foote, K. Sparck Jones and S. J. Young. Video Mail Retrieval: The Effect of Word Spotting Accuracy on Precision. ICASSP, pp309-312. IEEE, 1995.
[5] U. Glavitsch, P. Schäuble. A System for Retrieving Speech Documents. Proc. SIGIR, pp168-176. ACM, 1992.
[6] P. Schäuble, U. Glavitsch. Assessing the Retrieval Effectiveness of a Speech Retrieval System by Simulating Recognition Errors. Proc. HLT, pp370-372. ARPA, 1994.
[7] M. Wechsler, P. Schäuble. Indexing Methods for a Speech Retrieval System. In Proceedings of MIRO Workshop, Glasgow, September 1995. Available on the WWW at http://www-ir.inf.ethz.ch/ISIR-Papers.html
[8] BBC Television and Radio Schedules. Available on the WWW at http://www.bbcnc.org.uk/bbctv/
[9] M. F. Porter. An Algorithm for Suffix Stripping. In Program, 14(3):130-137, July 1980.
[10] G. Salton and C. Buckley. Term-weighting Approaches in Automatic Text Retrieval. In Information Processing And Management, 24(5), pp513-523, 1988.
[11] S. J. Young, P. C. Woodland and W. J. Byrne, HTK: Hidden Markov Model Toolkit V1.5. Entropic Research Laboratories Inc., Washington, 1993.
[12] D. A. James and S. J. Young. A Fast Lattice-Based Approach to Vocabulary Independent Wordspotting. In Proc ICASSP, pp I-377-380. IEEE, Adelaide, 1994.
[13] R. C. Rose and D. B. Paul. A Hidden Markov Model based Keyword Recognition System. In Proc. ICASSP, pp129-132. IEEE, Albuquerque, 1990.
[14] P. C. Woodland, C. J. Leggetter, J. J. Odell, V. Valtchev and S. J. Young. The 1994 HTK Large Vocabulary Speech Recognition System. ICASSP, pp I-73-76. IEEE, 1995.
[15] Bowker Saur Ltd. The Independent on CD-ROM: 1 October 1989 - 31 December 1990. London, 1991.
[16] S. M. Katz. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recogniser. In IEEE Trans. ASSP-35(3):400-401, 1987.
[17] J. Kupiec, D. Kimber and V. Balasubramanian. Speech-based Retrieval using Semantic Co-occurrence Filtering. In Proc. HLT, pp373-377. ARPA, 1994.
[18] S. Robertson and K. Sparck Jones. Relevance Weighting of Search Terms. In Journal ASIS, 27:129-146, 1976.