Technical Report

UCAM-CL-TR-430
ISSN 1476-2986

Number 430

UNIVERSITY OF
CAMBRIDGE

Computer Laboratory

Video mail retrieval using voice: Report on topic spotting

(Deliverable report on VMR task no. 6)

G.J.F. Jones, J.T. Foote, K. Sparck Jones,
S.J. Young

July 1997

15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
phone +44 1223 763500

https://www.cl.cam.ac.uk/

Technical reports published by the University of Cambridge
Computer Laboratory are freely available via the Internet:

https://www.cl.cam.ac.uk/techreports/

ISSN 1476-2986

Video Mail Retrieval Using Voice : Report on Topic Spotting (Deliverable Report on VMR Task No 6) *

G.J.F. Jones[†‡], J.T. Foote[‡], K. Sparck Jones[†] and S.J. Young[‡]

[†]Computer Laboratory, University of Cambridge,
New Museums Site, Pembroke Street
Cambridge CB2 3QG

[‡]Engineering Department, University of Cambridge,
Trumpington Street,
Cambridge CB2 1PZ

July 1997

Abstract

This report describes research on topic spotting in audio document retrieval carried out in years 2 and 3 of the Cambridge Video Mail Retrieval (VMR) project. Topic spotting within VMR was concerned with *ad hoc* querying of a message archive using classical information retrieval techniques developed from experience with text archives. The report describes experiments using three approaches to document indexing: fixed-vocabulary keyword spotting, open-vocabulary search term indexing using phone lattices, and message transcription using large vocabulary speech recognition. Additional experiments investigate the combination of these techniques for improved retrieval effectiveness.

*This project is supported by the UK DTI Grant IED4/1/5804 and SERC (now EPSRC) Grant GR/H87629. Olivetti Research Limited (ORL) is an industrial partner of the VMR project.

Contents

List of Tables

List of Figures

1 Introduction

This report describes the motivation, methods, experimental design and results, and evaluation of *topic spotting* in the Cambridge Video Mail Retrieval (VMR) project [Hopper et al., 1993]. The overall project objective was the development and evaluation of a prototype video mail retrieval system at ORL, Cambridge. The topic spotting investigation formed Task 6 of the overall VMR project plan. This report brings together all the relevant material for this major task, including both previously published and unpublished test results, to allow a comprehensive view of the whole.

In the VMR project topic spotting is taken to mean searching a spoken message archive for items hopefully about the same topic as a user's text request, and hence relevant to the user's information need.

Topic spotting in VMR is the same as ordinary *ad hoc* text document retrieval [van Rijsbergen, 1979]. A topic is defined within VMR to be a (complex) concept that may be linguistically expressed in different ways. The critical point about topic spotting is that there is no guarantee that if the terms in a search query and document text match that they are actually about the same topic (or indeed if they fail to match on words, that they are not about the same topic). We contrast this interpretation of the phrase with that sometimes encountered in speech research, where it refers to the use of *standing* topic definitions for such purposes as document routing or categorisation [Rose, 1991], [Wright et al., 1995] [Peskin et al., 1996].

VMR topic spotting experiments were carried out using several approaches to speech recognition for document indexing. Initial work concerned word spotting using a small fixed keyword vocabulary of 35 words with both speaker-dependent and speaker-independent recognition. At this point only the fixed keyword vocabulary could be used for searching. Work in the latter stages of the project concentrated on open-vocabulary word spotting via subword phone lattices and a limited investigation into the use of large vocabulary speech recognition. Phone lattice spotting and large vocabulary recognition enable a large (and potentially unlimited) search vocabulary to be used.

Topic spotting is thus the prime retrieval task both from the specific VMR application point of view and in the wider context of spoken document retrieval. We will henceforward use conventional information retrieval (IR) terminology for it, referring to user *requests* containing words, and to search *queries* consisting of index *terms* derived from these requests.

Specifically, as far as the general approach to indexing and retrieval we have followed is concerned, we have followed mainstream practice in assuming:

1. that the topics in which users are interested are adequately represented by sets of words (in fact normally word stems);

2. that the extent to which query terms are jointly found in a document is a fair indicator of document relevance to user need (so the output from a search is ranked by matching score); but

3. that this simple coincidence matching on terms alone is modified to take account of term weights, so query-document scores are sums of shared term weights not just counts of common terms.

Further, we have applied the well-established Robertson/Sparck Jones model of term weighting for this [Robertson and Spärck Jones, 1976] [Robertson and Spärck Jones, 1994]. Thus as far as retrieval is concerned our methods are conventional, and our work on topic spotting has concentrated on the issues of recognising query terms in spoken documents.

1.1 Experimental Strategy

Overall, the topic-spotting tests fit into a primary characterisation of retrieval conditions for the VMR experiments as follows:

	Vocabulary	Speakers	Terms
Stage 1	closed	closed	(keywords)
Stage 2	closed	open	(keywords)
Stage 3a	open	closed	(topics)
Stage 3b	open	open	(topics)

The experiments cover both the *speaker-dependent* (closed) case and *speaker-independent* (open) case. Better speech recognition performance for the former means retrieval performance can be expected to be higher than for the independent case, but is less realistic than the speaker-independent situation most likely to arise in practice.

In relation to vocabulary, there are further alternatives as follows.

1. In relation to individual words, whether:
   (a) acoustic processing in speech recognition uses whole or subword modelling;
   (b) indexing (and hence document matching) uses full words or stems.

2. In relation to the word set we define:
   (a) Word Spotting (WS): the use of a fixed *a priori* keyword list, the terms available for searching are well chosen for the domain in advance of speech recognition, but do not necessarily cover all the user's request words.
   (b) Large Vocabulary Recognition (LVR): the use of a large vocabulary system, e.g. covering 20,000 words. This extends the range of user request words that may be searched as index terms to all words in the finite pre-selected LVR vocabulary, though this does not necessarily capture all of the search terms. This issue of search terms missing from the LVR vocabulary particularly affects many proper names which cannot be anticipated in advance. Search terms and words spoken but not in the vocabulary are referred to as out-of-vocabulary (OOV) words.
   (c) Phone Lattice Scanning (PLS): the use of phone lattices which may be used to search for any word. These may be used as an indexing strategy in their own right, or alternatively to search for any request words which are OOV with respect to fixed vocabulary recognition systems.

This report reviews experiments from all stages of the VMR project, although since stage 1 and 2 are described in detail elsewhere [Spärck Jones et al., 1995, Jones et al., 1995a], while stages 3 and 4 represent the technology in the final system this report concentrates on this latter work.

1.2 Index Source Combinations

Our experiments also cover a range of subsidiary alternatives where different sources of indexing information can be variously combined to realise useful overall improvements in retrieval performance [Belkin et al., 1995]. The following combinations were investigated:

(a)	Speaker–Dependent Word Spotting	Speaker–Independent Large Vocabulary Recognition
(b)	Speaker–Independent Word Spotting	Speaker–Independent Large Vocabulary Recognition
(c)	Speaker–Dependent Phone Lattice Scanning	Speaker–Independent Large Vocabulary Recognition
(d)	Speaker–Independent Phone Lattice Scanning	Speaker–Independent Large Vocabulary Recognition

The various combined systems are progressively less constrained. (a) requires the choice of domain specific keywords, and speaker dependent acoustic models, while (d) has an open search vocabulary and speaker-independent modelling.

1.3 Comparison of Indexing Methods

From the speech recognition point of view, the key differences between WS, LVR and PLS are:

1. WS, processing looks only for occurrences of keywords without regard to their context, all speech other than that deemed to match a keyword is simply treated as a sequence of *filler* models (typically subword units, actually *phones* in our work). WS is prone to both *false alarms* (hypothesis of a search word where none exists) and *misses* (failure to hypothesise a search word which is actually present in the speech) in word matching. In the design of a WS system there is a trade off between the number of false alarms and misses. A stochastic acoustic match score threshold can be applied to remove the majority of false alarms, but this is also likely to exclude some proper matches.

2. LVR, generates a complete (though errorful) transcription of the document. The acoustic models are operated in conjunction with a language model. Due to the large available recognition vocabulary and the contextual information, false alarms are relatively rare; however misses are much more likely in LVR systems. This can occur for example where a work is relatively "unlikely" to appear in a certain context as measured by the language model and a more linguistically "likely", often acoustically similar, alternative may be substituted. These subjects are discussed in more detail in the VMR deliverable report on LVR [Jones et al., 1996b].

3. PLS, any word for which a subword phone structure is available can be searched for. The phone structure is typically contained in a dictionary which maps lexical entries to corresponding phone sequence(s). This is hence much more versatile than WS, however, processing is still intrinsically unreliable. This unreliability is partially due to the inaccuracy of phone recognition due to their acoustic confusability, typically

> only about 60% can be recognised correctly. Thus in practice it is found that a relatively deep phone lattice is required to capture a high proportion of the correct phones. The consequence of this is that many false alarms will arise in the scanning of the lattice. As for WS many of these can be removed by thresholding. Phone lattice spotting is described in more detail in the VMR deliverable report on PLS [Foote et al., 1996] and further in [Foote et al., 1997].

Ideally, from the retrieval point of view, and seeking the well-known advantages of redundancy, it would be most sensible to use all of WS, LVR and PLS. However in practice it is often not possible to generate a specific keyword list for a document archive and WS will not be available. Hence it is more important to assess retrieval using PLS or LVR either alone or in combination. The importance of redundancy in this type of environment is discussed in Section 4.1.2.

There is a further condition relevant to spoken document retrieval, namely the nature of the recording environment. A head-mounted microphone gives noticeably better results than a desk one, though the latter is more likely to be found in practice. We give results for both. It should be noted, however, that our documents were recorded in a noise-free environment, implying better recognition performance than would be achievable in many practical situations.

The material in the rest of this report is organised as follows: in Section 2 we describe our experimental data set and in Section 3 the speech processing employed in this work; Section 4 describes the information retrieval methodology employed in our work; we then present retrieval results for the systems in Section 5; finally, in Section 6 we summarise and comment on the main findings to be drawn from these sets of detailed runs.

1.4 Retrieval Experiments

Retrieval performance for spoken documents is compared with that for manual text transcriptions which are taken to define the standard of performance attainable for the data set.

Additionally we have used alternative 'phonetic text' transcriptions which are in principle a fairer base for comparison for WS and PLS. Phonetic text is generated by decomposing the complete word level manual transcription to the sub-word phone level. This phone sequence is then scanned for phone sequences matching search terms. Phone strings which match can then be counted as correct hits, even if as the word level they are semantically unrelated to the search term. The recognition argument here being that if the phone string is the same as the search word it is reasonable for the WS to hypothesise the word at this point. Thus, for example, when searching for the word *locate* we have observed its hypothesis during the utterance of *hello Kate*, the latter section of which consists of the phone sequence locate. Thus, many false alarms arise where unrelated acoustic events will often resemble valid words. When measured against the phonetic text standard hypothesis of locate at this point is thus judged to be correct [Jones et al., 1995b, Jones et al., 1995a] The difference between the two forms of transcription is small, and for simplicity in this report we refer only to the full text comparisons.

2 Data Provision

Experimental results reported here are for collections VMR1a and VMR1b. VMR1 is a specially constructed set of 300 messages described fully in [Jones et al., 1994], collections VMR1a and VMR1b consist of two separate request/relevance assessment sets. They are both overviewed in [Jones et al., 1995a] and a full description of VMR1b is contained in [Jones et al., 1996a].

2.1 The VMR message corpus (VMR1)

Because there was no available video mail corpus and existing speech corpora were not suited to retrieval experiments, an archive of messages with known audio and information characteristics was created in order to evaluate both speech recognition and message retrieval performance. Having to construct rather than select a test collection is regrettable but was unavoidable, and the test collection shares important properties with real test sets, e.g. variable topic overlap between documents.

Ten broad subject categories were chosen to reflect the anticipated messages of ORL video mail users, including, for example, "management" and "equipment." For the initial domain-dependent indexing using small vocabulary WS, a fixed set of 35 keywords was provided for the ten categories; thus the keywords "staff," "time," and "meeting" refer to the "management" category (though keyword-category assignment is not exclusive). The keyword set includes 11 difficult-to-recognise-correctly monosyllabic words (e.g "date" and "mail"), as well as overlapping words (e.g. "word" and "keyword") and word variants (e.g. "locate" and "location").

Fifteen speakers (11 men and 4 women) each provided about 45 minutes of speech data. This produced a total of 5 hours of read training data and 5 hours of spontaneous speech messages. The acoustic training data was used to create the speaker-dependent (SD) models for the recognisers and consisted of isolated keywords, read sentences containing keywords in context, and phonetically-rich sentences not containing keywords. For the message data, each speaker provided 20 spontaneous speech messages in response to 5 prompts chosen from 4 categories. The resulting 300 messages, along with their manual text transcriptions, served as a test corpus for the retrieval experiments presented later. The messages, though prompted, are fully spontaneous and contain a large number of disfluencies such as "um" and "ah," partially uttered words and false starts, laughter, sentence fragments, and informalities and slang ("'fraid" and "whizzo"). The messages were fully transcribed by hand, including non-speech events such as lip smacks, hesitations, and disfluencies. Basic punctuation was also added for ease of reading. These full transcriptions were used to evaluate both speech recognition and retrieval performance.

Data was recorded at a 16 kHz sampling rate, from a Sennheiser HMD 414 head-mounted microphone and the Medusa system desk-mounted microphone used in the ORL video mail system. For speech model training and recognition, the acoustic data was parameterized into a spectral representation at a 100 Hz frame rate.

The VMR1 message set is very small by text retrieval standards, but as an experimental corpus for spoken document retrieval it compares respectably with [Wechsler and Schäuble, 1995, McDonough et al., 1994], and is also comparable with speech processing test data used until recently for ARPA experiments [Young et al., 1994].

2.2 WSJCAM0

Speaker-independent acoustic models were trained using the WSJCAM0 British English spoken corpus. This consists of spoken sentences taken from the Wall Street Journal (WSJ). Data was collected for 100 British English speakers with equal numbers of male and female speakers drawn from a variety of age groups and regional backgrounds. The corpus contains a total of around 12 hours of spoken data. WSJCAM0 was collected at Cambridge University Engineering Department and further details are contained in [Robinson et al., 1995].

2.3 Retrieval Collection VMR1a

In order to obtain some test requests quickly, given a lack of users, we decided simply to exploit the prompts used to obtain messages in the database recording. Elements of the standard van Rijsbergen stop word list [van Rijsbergen, 1979] were deleted and to reduce variations in word form, the remaining query words were suffix stripped to stems using the standard Porter algorithm [Porter, 1980]. Thus we obtained a total set of 50 requests with corresponding simple term list queries. The average length of requests was 38.1 words, after removing the stop words this was reduced to 19.0 words. Including only suffix stripped terms in the 20K vocabulary derived from the Wall Street Journal (WSJ) used in our LVR tests there were on average a total of 18.2 terms per query. For the fixed keywords there were an average of 5.7 terms per query.

To obtain relevance assessments, the 6 recorded messages generated in response to each prompt were assumed relevant to the query constructed from that prompt. The 24 other messages in the same category, which are quite likely to contain similar words since they are closely related, are assumed to be not relevant. This whole procedure was somewhat crude, but we believe it gave us adequate material, from a term distribution point of view, for fair experiments.

Av. No of Words per Request	38.1
Av. No of Words per Request after removing van Rijsbergen stop list.	19.0
Av. No of Words per Request after removing words not in WSJ 20K word vocabulary	18.2
Av. No of Fixed Keywords per Request	5.7
No of Relevant Documents per Request	6.0

Summary of Collection VMR1a

2.4 Retrieval Collection VMR1b

We subsequently obtained a second set of more realistic requests and relevance assessments from the user community that supplied the database messages. A total of 50 requests was collected, 5 for each of the 10 categories defined previously. These were gathered from 10 users who each generated 5 requests and corresponding relevance assessments. This was achieved by forming 10 unique sets of 5 categories and assigning each to a user knowledgeable about the subject matter of the categories in the set. For each category requests and relevance assessments were generated as follows.

Requests Subjects were asked to form a natural language request based on the information given in a text prompt. One such prompt was formed for each of the message categories, described earlier, by combining the information given in the 5 message scenario prompts associated with the category. Hence, there were 10 prompts in total. Subjects were asked that their request include at least one of the fixed keyword associated with the category, as defined for the message collection phase.

Relevance Assessment Each request was converted to a query using the same method as VMR1a. The query was used to score each document transcription in the message archive using the standard query-document matching technique described in Section 4.1 using collection frequency weighting. The messages were then ranked in order by decreasing matching score.

Ideally users should assess the relevance of all messages in the archive to their request; however, even with the 300 document archive this was considered impractical. Thus a suitable message subset for assessment must be generated. For VMR1b the list of messages for assessment was formed by combining the 30 messages generated for the category to which the original message prompt belonged together with the highest scoring 5 messages which were not associated with the category. Thus the user was presented with 35 messages to assess for each request. To avoid sequence effects the order of presentation of these messages to the user was randomised. The subjects were presented with the transcription of each potentially relevant message and asked to marked it as "relevant", "partially relevant", or "not relevant". A full description of the formation of the Collection 1b naturalistic request set is contained in [Jones et al., 1996a].

Av. No of Words per Request	20.0
Av. No of Words per Request after removing van Rijsbergen stop list.	7.4
Av. No of Words per Request after removing words not in WSJ 20K word vocabulary	6.6
Av. No of Fixed Keywords per Request	2.6
Av. No of Highly Relevant Documents per Request	10.8
Av. No of Highly or Partially Relevant Documents per Request	17.2

Summary of Collection VMR1b

Apart from the greater realism, the main differences between VMR1a and VMR1b were that there were far fewer terms per query for the latter. The requests averaged 12.0 words. After removing the stop words this reduced to an average of only 7.4 content terms. On average 6.6 of these terms are found in the 20K vocabulary. The keyword-only versions of these requests contain an average of 2.6 terms. Compared to VMR1a there was naturally variation in the number of relevant documents per query as well as, in fact, a larger average number, 10.8 highly relevant and 17.2 highly or partially relevant. All retrieval results in this report for VMR1b use only the highly relevant document collection.

3 Speech Recognition Techniques

All the speech recognition work in the VMR project is based on Hidden Markov Models (HMMs). The HMMs exploited in the WS, LVR and PLS systems used *acoustic models* of individual phones, and also (for LVR) a *language model* to capture word associations. The speech recognition systems for the work reported here used the HTK tool set developed at Cambridge University Engineering Department [Young et al., 1993]. All the recognition techniques described below deliver an *acoustic score*, the log-likelihood that the observed sound or sound sequence is actually an instance of the matching phone or word model.

3.1 Fixed Vocabulary Word Spotting (WS)

The basic idea of speech recognition using fixed vocabulary WS is to correctly identify occurrences of predefined keywords while mapping all other acoustics events to a general background subword filler model or silence. Although described in detail elsewhere [Foote et al., 1994, Jones et al., 1995b, Foote et al., 1995], a short outline of our WS system for fixed keywords is given here for completeness.

As described previously, two types of acoustic models were investigated here: speaker-dependent models and speaker-independent models.

Speaker-dependent modelling For speaker-dependent (SD) WS separate whole-word keyword models and phone filler models were built for each of the 15 speakers. Each keyword model was trained with the 10 occurrences of the word in the VMR1 training data. SD filler phone models to represent non-keyword speech were trained on the remaining speech data. The filler models here were all monophones, phone models which are independent of context. All examples of these phones occurring in the filler training data were used to train a single HMM model of each phone.

Speaker-independent modelling In the speaker-independent (SI) WS system, each keyword is modelled by concatenating the appropriate sequence of subword phone models (obtained from a phonetic dictionary). Phones vary depending on acoustic context and, as will be demonstrated for phone-lattice spotting, using context-dependent phone models can improve recognition performance. Biphones, which take into account the previous or next phone context, were used at the beginning and end of keywords; while triphones, which take into account both previous and next phones, model their internal structure. For example, the keyword "find" is represented by the model sequence `f+ay f-ay+n ay-n+d n-d`. Keyword models were constructed from a set of 8-mixture word-internal tied-state triphone HMMs trained on the WSJCAM0 British English speech corpus [Robinson et al., 1995] using a tree-based state clustering technique [Young et al., 1994]. Non-keyword speech is modelled by a SI filler model consisting of an unconstrained parallel network of monophones. Thus all speech is recognised as either a keyword or a phone from the filler network.

Word Spotting Procedure WS is done with a two-pass recognition procedure. First, Viterbi decoding is performed on a network of just the filler models. This yields a time-aligned sequence of the maximum-likelihood filler monophones and their associated log-

	Data set	
	Head (%)	Desk (%)
Speaker dependent	81.2	76.4
Speaker independent	69.9	55.9

Table 1: Average Figures of Merit for SD and SI WS.

	Data set	
	Head (%)	Desk (%)
R13	77.1	57.8
R75	80.5	65.5

Table 2: Average Figures of Merit for SI WS with Speaker Adaptation.

likelihood scores. Second, another Viterbi decoding pass is done using a network of the keywords, silence, and filler models in parallel. In a manner similar to [Rose, 1991], keywords are rescored by normalising each hypothesis score by the average filler model score over the keyword interval. Normalisation helps ensure that true keyword hits have scores greater than false alarms. Because low-scoring words are more likely to be false alarms, the operating point of the recognition system may be adjusted by ignoring words with a score below a given threshold.

The accuracy of a word spotter thus depends on its threshold and cannot be expressed as a single number if false alarms are taken into account. An accepted figure-of-merit (FOM) for word spotting is defined as the average percentage of correctly detected words as the threshold is varied from one to ten false alarms per word per hour. (This is quite similar to retrieval average precision, where precision is averaged as output is varied.)

The VMR corpus is realistic in that it contains speakers with varied backgrounds and accents. This is not significant for SD modelling where separate models are built for each speaker. However, the SI models are trained from exclusively using British English spoken data. In this case recognition performance for the north American speaker in the VMR1 archive is noticeably degraded.

Table 1 summarises FOM results for SD and SI WS using both head and desk microphones. Observations from this table are that SD modelling is better than SI, and that the head-microphone gives better recognition performance than the desk-microphone. Both of these results are expected and the overall performance is in line with that found by other researchers using this approach to WS.

Speaker Adaptation In an attempt to ameliorate the problems associated with non-native speakers, and increase word spotting performance in general, speaker-independent acoustic models may be adapted to individual speakers as described in [Foote et al., 1995]. In this procedure a small amount of "adaptation" data is used to generate a personal modified HMM model set which better represents the speech of the individual speaker. The approach chosen was maximum-likelihoods linear regression because it has been shown to improve recognition with a comparatively small amount of adaptation data [Leggetter and Woodland, 1995]. This method involves adapting only some of the HMM parameters (the means of the HMM Gaussian mixtures) to increase the likelihood of the adaptation data given the models. Varying amounts of the VMR1 training corpus were used as enrollment data for speaker-adaptation experiments.

WS performance using speaker adaptation is shown in Table 2. The *R13* row used 13 utterances of enrollment data containing in all 2 occurrences of each keyword. The *R75* row used the full 75 "r" sentences from the VMR1 training material, containing 5 utterances of each keyword. A more detailed examination of these results shows that adaptation does not uniformly improve performance for all speakers. However the average increase is substantial, and is particularly dramatic for our American English speaker. As shown in Table 2 using a small amount of enrollment data improved the FOM performance substantially for the head microphone, although much less for the desk microphone. A greater amount of adaptation data produced further improvement in the FOM of the head-microphone, but proportionally a much larger improvement for the desk microphone.

This type of speaker adaptation is referred to as *supervised* since the correct transcription of the enrollment data is known by the recogniser. In operation this requires the sender to speak some given enrollment text so that models can be suitably adapted in advance of message recognition. In practice, this may not be possible for the VMR system since it is quite probable that there will be no opportunity to gather the enrollment material for messages from a new speaker. An obvious extension to this work would be to investigate *unsupervised* adaptation where the parameters are modified without use of *a priori* transcriptions.

3.2 Phone Lattice-based Word Spotting (PLS)

Data set			
Head (%)			Desk (%)
SD mo	SI mo	SI bi	SI bi
73.6	48.0	60.4	51.5

Table 3: Average Figure of Merit for PLS.

The PLS word spotting technique involves searching a phone lattice for the sequence of phones corresponding to a particular search term [James and Young, 1994]. A phone lattice is a directed acyclic graph whose nodes consist of start/end times, and whose arcs are putative phone occurrences, which are labelled with the phone's acoustic score. Phone lattices may be computed in advance, and rapidly scanned for an arbitrary phone sequence at search time. In the VMR system this search is rapid enough to allow interactive scanning of search word input at retrieval time.

For the head microphone experiments reported here, separate phone lattices were generated using three model sets: 8-mixture SD monophones, 8-mixture SI monophones, and 8-mixture SI biphones. In addition to the acoustic models, bigram phone transition probabilities were enforced in a null-grammar network. Phone transition probabilities were trained using the transcriptions of the "z" data taken from VMR1 and are therefore independent of the VMR experimental domain. A phone lattice is generated from the n-best paths through a HMM network, given the network, models, and (unknown) acoustic data. The value of n controls the number of simultaneous hypotheses that may end at a given time; thus n controls the average lattice depth, and the number of possible paths through it. The choice of n represents a trade off between: correct hits, false alarms, misses, and the overall size of the lattice. This latter factor is important since in order to provide very rapid scanning the lattices must be available online in RAM; of course, the smaller the lattices actually are, the easier this is to achieve.

At search time, the phone lattice is scanned for the phone sequence corresponding to the query search word. The phonetic composition of a search word is derived from a dictionary (British English pronunciations are taken from the Oxford Learner's Dictionary). Once a phone sequence is found (corresponding to a putative occurrence of the search term), an acoustic score for the term is estimated from the scores of the component phones using a similar procedure to that used for WS. In general, phone-lattice spotting accuracy is much poorer than the fixed-keyword spotting described earlier. For each model set the FOM for PLS was computed for the same 35 keywords as used for the standard WS system. FOM values for the PLS systems are shown in Table 3. The head-microphone data results for monophone models indicate that, as for WS, SD modelling gives improved performance over SI modelling. Although SI performance can be improved through the use of the more detailed biphone models. (There was insufficient acoustic training data to investigate the performance of SD biphones.) Since they are clearly the best available practical models, only SI biphone models were built for the desk microphone data. As for WS, performance with the desk microphone data is somewhat degraded compared to head microphone data, although not disastrously so. We would expect to observe similar performance trends with SD and SI monophones to those observed for the head microphone models. The desk microphone SI biphones were used in the final VMR demonstration system.

Whilst overall the PLS figures are lower than those for the WS system, it is important to remember that the phone lattices are completely general and that any set of words can be searched for without further speech recognition effort.

3.3 Large Vocabulary Recognition (LVR)

LVR is potentially the best available technique for speech recognition in retrieval applications. LVR attempts to provide a complete transcription of the message. This has the advantage that a complete inverted file can be constructed in advance of retrieval with all terms available meaning that multi-pass retrieval strategies such as relevance feedback might be employed. Also, an automated text transcription has the advantage over a phone lattice of being very compact. Large vocabulary continuous recognition is only now becoming practical [Young et al., 1994] and only limited evaluation of its potential has been possible in the VMR project. LVR will wrongly transcribe a significant number of document words. Many of these recognition problems arise due to OOV effects. For example, when an OOV term is present in the message to be transcribed an alternative word (or words) which is (are) in the vocabulary must be transcribed into this position. This recognition error will often introduce other errors nearby due to poor word boundary alignment and probabilistic combination between the acoustic and language models. However despite the current shortcomings of LVR technology, it has already demonstrated its potential utility for retrieval as described here and in more detail in [Jones et al., 1996b].

For the LVR experiments, a set of 8-mixture cross-word triphones was trained on the WSJCAM0 British English speech corpus, of read *Wall Street Journal* material [Robinson et al., 1995]. Ideally a suitable language model would be built using a large transcription archive of material typical of the application domain. Unfortunately since there was no available archive of this type, the standard WSJ 20K bigram language model from MIT Lincoln Labs was used. The WSJ triphone set and bigram language model when taken together yielded a 53% word recognition accuracy rate. This is low compared to read speech, where accuracy rates can exceed 90% in a limited domain, but is respectable given the difficulty of the spontaneous VMR task and the domain mismatch. Many factors impact recognition performance adversely. For example, the VMR1 corpus has a significant out-of-vocabulary rate of 3.15%, including 4 of the 35 frequently-occurring fixed keywords, and the WSJ North American business news language model is highly inappropriate for informal British English monologues. Also problematic is the exclusively read training data, the spontaneous nature of our test speech, the lack of disfluency modelling for it, and its non-uniform accents (British, American, and Middle European) [Jeanrenaud et al., 1995]. However, as is shown later in this report, even the imperfect recognition of the existing system results in respectable retrieval performance.

3.4 Thresholding

As outlined in preceeding sections, acoustic word spotting using either WS or PLS is prone both to missed words and to false alarms when seeking search terms. In the case of WS the filler model may be incorrectly deemed more likely than the search word or vice versa. In PLS the uncertainly and reduced search space of the lattice may also lead to the presence of phone strings which correspond incorrectly to search terms causing false alarms, or one or more of the correct phones may be missing from the lattice leading to a miss.

Since many false alarms will have a lower score than true hits, a threshold is normally set on the acoustic score. Words with scores above the threshold are considered true hits, while those with scores below are considered false alarms and ignored. Choosing the appropriate threshold is a tradeoff between the number of Type I (missed words) and Type II (false alarm) errors, with the usual problem that reducing one increases the other. Retrieval performance varies with the choice of score threshold. At low threshold values, performance is somewhat impaired by a high proportion of false alarms (Type II errors); conversely, higher thresholds remove a significant number of true hits (Type I errors), also degrading performance. In our work thresholding was applied to both WS and PLS output hypotheses, but it is not required for LVR because there are few false alarms.

4 Information Retrieval Techniques

For our experiments, standard indexing and matching techniques were applied both to the text transcription files, and to the quasi-transcriptions and other search word hypotheses generated by the speech recognition engines. Performance for the text transcriptions could thus be used as a reference standard for the various speech retrieval strategies.

4.1 Indexing Methodology

The indexing methodology follows standard text indexing and matching techniques. The text transcriptions and LVR output files were first processed to remove stop words. Several stop word lists were investigated but most consistent results were obtained using the standard van Rijsbergen list [van Rijsbergen, 1979]. Results in Section 5.2.1 show retrieval performance for a number of stop word lists. (Stop word removal is not required for the fixed keyword WS and PLS since they only return putative hits from the fixed vocabulary or the current query respectively and so contain no stop words.) Next all query terms and hypothesised document contents from all sources were suffix stripped using the Porter algorithm [Porter, 1980]: see the next section for a discussion of stemming in relation to speech data.

Retrieval tests compared *unweighted uw* matching performance with two forms of weighting. These were the standard *collection frequency weight cfw* (also called inverse document frequency weight), and the *combined weight cw* that incorporates within-document term frequencies and is normalised for document length (defined in [Robertson and Spärck Jones, 1994] and derived in [Robertson and Walker, 1994]; the *cw* scheme reflects the City University work for TREC [Robertson et al., 1995]). The *cw* weight for each term in each document is calculated as follows:

$$cw(i,j) = \frac{cfw(i) \times tf(i,j) \times (K+1)}{K \times ndl(j) + tf(i,j)}$$

where *cw* (i,j) represents the *cw* weight of term i in document j, $tf(i,j)$ is the document term frequency and $ndl(j)$ the normalised document length. $ndl(j)$ is calculated as

$$ndl(j) = \frac{dl(j)}{\text{Average } dl \text{ for all documents}},$$

where $dl(j)$ is the total length of j. The combined weight constant K has to be tuned empirically: after informal testing a value $K = 1$ was selected.

4.1.1 Document Length *dl(j)*

The document length is ordinarily measured as the number of term occurrences in the document. This measure of $dl(j)$ is suitable for text and LVR where full transcriptions are available. For WS the document length is the number of terms from the fixed keyword vocabulary found in the document and in PLS the document is represented only by the search terms found for the current query; but the latter in particular may not be a good representation of the document length, i.e. long documents may appear short if they are not rich in the search terms of a particular query. However since $ndl(j)$ is the ratio between different document lengths, the absolute length is not important and alternative measures of $dl(j)$ can be considered. In our PLS tests we examined two alternative measures of $dl(j)$:

- the number of phones found in the most likely phone path, which is easily computed using the Viterbi algorithm during the speech recognition phase. We reason that on average the number of phones in a document is representative of the number of words.

- the total length of the document in seconds.

4.1.2 Redundancy

Problems associated with word sense and category ambiguity are familiar in text retrieval. These problems are compounded in spoken documents since there is always some ambiguity about whether a putative word is actually present. Some ambiguities may be resolved in the speech case by different pronunciations, but more ambiguities arise from homophones and misrecognised word boundaries. However as with text retrieval, redundancy can be exploited to reduce these uncertainties and ambiguities. Redundant information is basically multiple explanations of the same idea. Within a document redundancy is information which doesn't necessarily add to message of the document, but may help to remove ambiguities of sense or uncertainties. Thus a file containing the words "blind" and "venetian" may be about "venetian blinds" or "blind venetians." However if the word "sunblind" appeared in the same document this uncertainy would be resolved. If all of "venetian" "blind" and "sunblind" appear in the query, although the last term is apparently unnecessary it will potentially assist in discriminating relevant documents from irrelevant ones. One statement alone is useful, both together will improve chances of getting the correct document. LVR is thus potentially important for spoken document retrieval, compared with WS, as it significantly increases the number of potential matching keys. It can nevertheless still miss term occurrences due to recognition errors and OOV terms. Phone lattice spotting can make more terms (such as proper names) available for matching although confidence in the accuracy of individual putative hits is lower. If a domain specific set of OOV terms is available, more reliable OOV term searching could also be carried out with a fixed vocabulary word spotter.

4.2 Word Stemming Issues

As outlined in Section 4.1, text retrieval attempts to overcome mismatches between different *word forms* by using suffix stripping. Common approaches such as the Porter algorithm [Porter, 1980] use heuristic rules, e.g. application of these rules to `locate` and `location` both produce the stem `locat-`. Matching between text query and document terms here is clearly on the letters of the terms. The location of terms in spoken documents requires rather more careful examination. Since we postulate that performance on text represents the standard by which spoken document retrieval is best judged, the objective in indexing spoken documents should be to produce document indexing identical to that observed for text. This is implicitly achieved with perfect automatic transcription. However, given that this is at present not achievable, word search strategies must be selected which most closely produce the same index information as an accurate transcription. Note, this objective is not necessarily the same as seeking the lowest word recognition error rate. For example, improved function word recognition performance which led to degradation in content word recognition may actually worsen retrieval performance. Degraded retrieval performance might occur despite overall improved word by word recognition performance since function words are of no benefit to retrieval and are hence usually removed using stop word lists whereas a useful content word may be misrecognised. Finally, it should be remembered that we are engaged in the development of an operational system and selected strategies should not be grossly inefficient in their requirements for computation or storage.

As has been outlined previously, all words are composed of fundamental acoustic phone units, e.g. locate is composed of the phone sequence l ow k ey t and location is likewise composed of the phones l ow k ey sh n. This has various implications for word searching using the different recognition strategies investigated in the VMR project. For clarity these are considered separately for the individual strategies.

4.2.1 Fixed Vocabulary WS

There are various possibilities here as candidates for the most effective indexing method. In detailing them we will make use of the informal notion of *base* form for words, as illustrated by the lead words used in conventional dictionary entries, for example verb infinitives or singular nouns. These lead words are often simple, compared with their related variant forms. They are also often pronounced in the same way as the *acoustic stem* of their associated variants, e.g. the base whole word locate sounds like the acoustic stem LOCAT-[1] in the word located. The possible indexing strategies are:

1. Look for all known forms of each keyword, for example we could search for each of locate, locates, located and locating individually. Of course if the variant set is simply informally provided by the user there may be other word forms in the file that are not anticipated in advance, but they will often form close acoustic matches with one of the supplied forms. This indeed applies not only to searcher-supplied variant sets but also to ones obtained in some relatively systematic way via a linguistic generation program or a reference dictionary. But whatever the source of the search word set, it is obviously computationally expensive and hence unattractive to look for very similar words in parallel.

2. Taking into consideration the computational inefficiency issue, we could look just for the given base form of the request keyword, e.g. locate, on the assumption that this will pick up other forms of the word with the same e.g. locates, or similar acoustic stem e.g. location. In evaluation we count only hypotheses of any word with the same common base as the given keyword as correct, as determined by linguistic checking e.g. dictionary reference. This should filter out false alarms with the same acoustic stem as the base keyword but no proper word form relationship e.g. (we imagine) occasion, which might appear in the signal to have the acoustic stem LOCAT-.

3. We proceed as with the previous methods, looking just for the given base form, but consider only exact matches as correct. This would produce the same WS output as 2, but different WS performance figures would result. In principle the same eventual retrieval output would be observed as in 2.

4. Given that the linguistic reference apparatus required for the two previous methods using bases may not be available, look for the suffix-stripped stem of the given request keyword. This suffers from the same limitations as the previous base-search technique, exacerbated by the fact that matching will often be more unreliable than for a full word. This is both because stems are ordinarily shorter than full words and because, as a stripped stem not be a linguistically-valid entity, the items it matches may be completely unrelated.

[1] Upper case here indicates acoustically derived stem as opposed to suffix stripped textual stem.

However even with these various limitations, searching for single items rather than variants is clearly attractive, so the question is which method among the last three is best in practice. Thus how, for example, does the easiest and cheapest approach, namely that using stems, perform? Investigations in text form of different words with the same suffix-stripped stem as one of the 35 VMR1 keywords revealed that 98.5% of those in a slightly-augmented standard BEEP phonetic dictionary [Robinson et al., 1995] had the same acoustic stem as the base keyword and further that this was also the acoustic form of the base keyword. For example, `locate`, `locates`, `located` and `locating` all have the same suffix-stripped stem `locat-` and all have the same acoustic stem, `LOCAT-` (phonetically `l ow k ey t`), as the whole-word base `locate`. We could thus search only for the word `locate` and be reasonably confident of identifying occurrences of the other word forms. They will of course only be identified by the WS as `locate`, but for retrieval purposes this is all that is required. The assumption that bases will match other word forms via their acoustic stems is only ever partially true, since strictly speaking there will be some slight modification of the final stem phone in the forms due to the slight changes of context. For example the end phone `t` for the base `locate` will be varied in the word endings `-TED` (phonetically `t ih d`) and `-TING` (phonetically `t ih ng`). This might result in some misses or low acoustic match scores (possibly resulting in removal during thresholding) when searching using only `locate`. However this phone variation is likely to be a marginal effect.

A more subtle effect is exhibited where the base form changes, for instance `locate` and `location` have the same suffix-stripped stem `locat-` but a different acoustic stem. The final phone `t` in `locate` is modified to `sh` in `location` which is then followed by the final phone `n`. This sort of phenomenon would seem to inhibit the likelihood of `locate` matching on occurrences of `location` if the latter were to be removed from the keyword list. In our study examples of these variations were comparatively rare, but the study was only on a small scale.

In our evaluations we adopted strategy 2, i.e. we used just the keyword base forms for searching. But we did not carry out a detailed analysis of the relative accuracy of matching only with base forms or with alternative word forms as well, or of the frequency of alternative word forms within the test data.

It should be noted that for test comparisons with *text* retrieval using keywords, the word normalisation usually implied by having a restricted indexing vocabulary must also be supplied. For our experiments this was done using standard Porter stemming. Thus in order to generate fixed keyword representations from the manual document transcriptions, each word in the transcription was suffix stripped and compared to the available suffix stripped keywords; suffix stripped document terms which matched one of the suffix stripped keywords were included in the keyword only representation of the document text. Fixed keyword queries for VMR1a and VMR1b were generated from the open vocabulary text form in the same way. Generally when keywords are referred to, they should be understood as involving word form normalisation.

4.2.2 PLS

The matching situation is similar for PLS. We can search the lattice using the same strategies as for fixed vocabulary WS. However, there is a slight further complication. In the case of fixed vocabulary WS we know that the search term is in the base form of the word and we believe that this will match the different word forms with reasonable accuracy. However, each search word form in PLS comes from a free text request where the user may submit any form of any word they choose. Thus in addition to the options for fixed vocabulary WS we can either:

1. look for the word in the form as it appears in the request, or
2. map this to some common base form(s) for all words which suffix strip to the same stem as the request word.

The former option is easier to implement and potentially more reliable in actually spotting the word form in the request correctly, since it will usually be longer. However, the assumption that this word form will map correctly onto other forms of the word would appear to be weaker here than for WS, since the end of the particular word form which appears in the request may not match well onto those of other word forms with the same acoustic stem.

The use of a consistent base form would obviously be problematic since a decision must be taken as to exactly what the correct phonetic root form should be and its phonetic analysis supplied. This would require additional manual preparation in the phonetic dictionary and is therefore not attractive in the design paradigm of the VMR system.

The final VMR demonstration system actually operates using the simpler first approach. As for fixed vocabulary WS no careful analysis of the accuracy of the assumptions made was attempted. In operation word hypotheses from PLS are stemmed using the Porter algorithm before being added to the inverted file. Calculation of the query-document matching score is thus at the level of stems. The effect being that words in a request with the same stem will appear as the same search terms in the inverted file, as would be the case for standard text retrieval matching.

An attractive way to search the lattice would be using a tree–clustered word modelling combining the features of the different word forms, so that all known possible instantiations could be searched for in a single pass over the lattice or possibly using a subset consisting of the request word and one or more general common root word forms. Shortage of time prevented this additional investigation from being carried out.

4.2.3 LVR

The objective in LVR is to make a complete transcription of the spoken data. The vocabulary may contain all, some or none of the different forms of a particular search word. In LVR the IR system cannot dictate the form of the search terms, but only has access to the pre-computed pseudo-transcription which is constrained to contain only the exact words in the LVR vocabulary. Thus term matching is possible only within these vocabulary constraints. In retrieval the LVR transcription is processed in the same manner as standard text. Thus stop words are removed and all remaining words are suffix stripped using the Porter algorithm.

Since the LVR transcription is dependent on not just acoustic matching, but also the language model, substitution of different word forms which are present in the vocabulary for those which are not or have been poorly articulated may be less practicable in LVR since the language model may score such substitutions very poorly. For example, the word `locate` may be unlikely to be substituted for the OOV `located` if the positioning of `locate` in this position is poorly scored in the language model. If the speaker said ... `it is located in the ...` the n-gram likelihood of ... `it is locate in the ...` is likely to be very low and another word sequence combined with the acoustic model may well score better, for example something like ... `this location is ...` might score better. From a retrieval perspective this would be perfectly acceptable, but none of the words has actually been recognised correctly, i.e. the absence of the word `located` from the vocabulary has led to several recognition errors rather than just one. As should be clear from the foregoing discussions, the presence of OOV words is inevitable in current LVR systems. Again, an exhaustive analysis of these effects in our LVR transcription output was not attempted for the VMR1 archive.

4.2.4 Summary of Adopted Word Stemming Strategies

As the details of the treatment of words are rather complex, and they differ for the various indexing techniques, the following table summarises the data.

		Available Indexing Vocabulary	When Term Added to Inverted File	Document Terms in Inverted File	Query Terms from Request	Matching Terms at Retrieval Time
Text		All Words	Index Time	stem	stem	stem
		Fixed Keywords	Index Time	stem	stem	stem
		All Words in 20K Dictionary	Index Time	stem	stem	stem
Spoken Docs	WS	Fixed Keywords	Index Time	stem	stem	stem
	PLS	Words as Instantiated in Request	Search Time	stem	stem	stem
	LVR	All Words in 20K Dictionary	Index Time	stem	stem	stem

This is the formal picture: however because, as already discussed, acoustic matching is only approximate, there may be some 'quasi'-base effects with PLS and LVR.

4.3 Index Combination Methods

It has been shown in text retrieval that combining multiple evidence sources can give overall improved retrieval performance, for example in [Belkin et al., 1995]'s comparative TREC-2 study. Belkin et al. considered two approaches to information combination, referred to as 'query combination' and 'data fusion'. In query combination, multiple queries for the same information need are merged into a single query, from which a single ranked output list is generated. In data fusion, multiple ranked output lists of documents (from different data representations) are combined to form a single overall ranked list. The methods described below use elements of both these techniques, tailored to suit the particular evidence conditions of spoken document retrieval.

4.3.1 Data Fusion

In our *data fusion* work, matching scores for documents that have been computed independently by different indexing systems, i.e. WS, PLS, LVR, are added to form a final composite score for each document. Since it is not clear whether scores for types of source are commensurable, we tried both normalising with respect to the highest scoring document in each list, and leaving scores as they were. With or without normalisation, the result is a single ranked list using the composite scores.

4.3.2 Data Merging

In our *data merging* strategies, evidence from different indexing sources is combined in a way analogous to Belkin et al's query combination. Specifically, word hypotheses from the indexing systems are merged to obtain a single term list for each document before computing the document's matching score. Hypotheses from the LVR output may be either augmented with all putative hits from a word spotter (PLS or WS), or only with those outside the WSJ 20K vocabulary. However, where all evidence from both sources is combined, search keys are counted twice if hypothesised at the same position by both systems. Searching for the same term using separate systems is not necessarily a drawback as it may help counteract acoustic stemming problems which may result in LVR misses when the term as instantiated is not in the LVR vocabulary (see Section 4.2).

Because they are frequency based, *cfw* weights may be affected by spurious keys in other documents due to PLS false alarms. *cw* weights which take into account within-document term frequencies, may also be adversely influenced by multiple counts of the same term, in addition to the effects of these existing false alarms. The overall length of a merged document is taken as the sum of all terms derived from both evidence sources for that document.

5 Retrieval Experiments

In this section we present our experiments in a series of comparisons as follows. First, in Section 5.1, Tables 4–23, we give a brief summary of experimental results using fixed vocabulary keyword spotting (WS), including results for text transcriptions and both SD and SI acoustic modelling. Following this in Section 5.2, Tables 24–43, more detailed retrieval results are given for open-vocabulary PLS retrieval. Again, initial open-vocabulary retrieval results are given for text transcription; these are used as the datum for all subsequent retrieval experiments. Next retrieval results are given for content-indexing using PLS spotting with various acoustic models. In Section 5.3, Tables44–65, we give results for large vocabulary LVR retrieval. The final results Section 5.4, Tables 66–97, cover retrieval using various data combination methods.

We have deliberately laid out the results in as full and regular a way as possible, so there are many tables. Thus we follow a pattern of giving results for Collection 1a and then Collection 1b for each particular performance factor. Some of these factors are *environment variables*, e.g. head or desk microphone recording, others are *system parameters*, e.g. choice of stop list. We have tried to follow a consistent ordering, with first head microphone and then desk microphone, first speaker-dependent (SD), then speaker-independent (SI), then speaker-adapted. More specific parameters apply to individual methods, e.g. monophone or biphone modelling for PLS. The individual tables also cover the various term weighting techniques, namely *uw*, *cfw* and *cw* (see Section 4).

Throughout this section where thresholding is used results are shown at the best *aposteriori* threshold. A detailed description of thresholding effects is given in [Spärck Jones et al., 1996]. All thresholds were chosen to maximise average precision, slightly different thresholds would often be chosen to maximise performance at a fixed cut off level such as used in [Spärck Jones et al., 1996]. Thresholding *aposteriori* gives a more favourable picture of performance than the *apriori* thresholding that would be necessary in practice, but this is not critical in the present context.

We recognise that with a small test collection such as our's specific figures are neither reliable nor significant: we concentrate therefore on the general picture that emerges from the results. For computing retrieval performance we show both document retrieval precision at rank cutoffs of 5, 10, 15 and 20 documents, and average precision, both computed using the standard TREC procedures[2].

5.1 Fixed Vocabulary Retrieval : WS Tests

Work described in this section is taken from Stages 1 and 2 of the VMR project and covers investigation into WS retrieval with the fixed keyword vocabulary. More detailed results and analyses appear elsewhere for SD in [Spärck Jones et al., 1996] and SI in [Jones et al., 1995a].

5.1.1 Text Retrieval

Tables 4 and 5 show manual text transcription retrieval results using only terms in the fixed keyword vocabulary. As will be noted for many other indexing techniques, the use of progressively more sophisticated weighting schemes produces improved retrieval performance.

5.1.2 Speaker-Dependent Modelling

Tables 6 and 7 show results for retrieval using SD head-microphone WS and Tables 8 and 9 show results for retrieval using SD desk-microphone WS. As anticipated performance is degraded somewhat with respect to text for both head- and desk-microphone modelling due to recognition errors. As would be anticipated from the WS results in Section 3.1, retrieval performance degradation is worse with the desk-microphone models. Encouragingly, retrieval performance for head-microphone modelling is still in excess of

[2] All retrieval performance figures are computed using the *trec_eval* software from TREC 2 developed (and kindly supplied to us) by Cornell University.

		Text		
Weighting Scheme		*uw*	*cfw*	*cw*
Precision	5 docs	0.264	0.296	0.300
	10 docs	0.222	0.250	0.270
	15 docs	0.192	0.213	0.236
	20 docs	0.170	0.193	0.208
Av Precision		0.293	0.332	0.358

Table 4: VMR1a retrieval precision for fixed keywords from manual text transcriptions.

		Text		
Weighting Scheme		*uw*	*cfw*	*cw*
Precision	5 docs	0.342	0.350	0.342
	10 docs	0.281	0.308	0.294
	15 docs	0.260	0.297	0.299
	20 docs	0.242	0.281	0.280
Av Precision		0.296	0.332	0.346

Table 5: VMR1b retrieval precision for fixed keywords from manual text transcriptions.

90% for both VMR1a and VMR1b; while that for desk-microphone modelling is still better than 80% in both cases.

5.1.3 Speaker-Independent Modelling

Tables 10 and 11 show retrieval results using SI head-microphone modelling WS, and Tables 12 and 13 show retrieval using SI desk-microphone modelling WS. For both head-microphone and desk-microphone modelling retrieval performance for SI modelling is degraded more than SD modelling with respect to text. This would be expected since, as noted previously in Section 3.1, the WS performance is not as good for SI modelling because of the generalisation of the acoustic models to any speaker.

As described in [Foote et al., 1995] there are a number of parameters in the WS which must be set empirically. The results shown here are the best obtained by a large and careful examination of the parameter space. Marginally better retrieval results might be obtained by a more exhaustive examination of the space. However, it is very unlikely that the improvement would be sufficient to be able to attach any statistical significance to differences between such results and those shown here.

5.1.4 Speaker Adaptation

As was shown previously in Section 3.1, speaker adaptation can be used to significantly improve WS performance for SI models. Tables 14 and 15 show retrieval performance for head-microphone models adapted using the R13 enrollment data (see Section 3.1), and Tables 16 and 17 show retrieval performance for head-microphone models adapted using the R75 enrollment data. It can be observed that improvement in retrieval performance is

Weighting Scheme		SD Head Models		
		uw	*cfw*	*cw*
Precision	5 docs	0.232	0.256	0.260
	10 docs	0.192	0.222	0.234
	15 docs	0.169	0.195	0.213
	20 docs	0.156	0.171	0.187
Av Precision		0.259	0.295	0.316

Table 6: VMR1a retrieval precision for WS using SD head-microphone modelling.

Weighting Scheme		SD Head Models		
		uw	*cfw*	*cw*
Precision	5 docs	0.333	0.350	0.333
	10 docs	0.265	0.321	0.296
	15 docs	0.238	0.285	0.289
	20 docs	0.207	0.260	0.266
Av Precision		0.265	0.312	0.330

Table 7: VMR1b retrieval precision for WS with SD head-microphone modelling.

Weighting Scheme		SD Desk Models		
		uw	*cfw*	*cw*
Precision	5 docs	0.244	0.260	0.272
	10 docs	0.182	0.214	0.238
	15 docs	0.167	0.191	0.210
	20 docs	0.142	0.166	0.177
Av Precision		0.241	0.283	0.299

Table 8: VMR1a retrieval precision for WS using SD desk-microphone modelling.

Weighting Scheme		SD Desk Models		
		uw	*cfw*	*cw*
Precision	5 docs	0.296	0.308	0.338
	10 docs	0.273	0.300	0.302
	15 docs	0.246	0.274	0.282
	20 docs	0.215	0.245	0.254
Av Precision		0.254	0.296	0.315

Table 9: VMR1b retrieval precision for WS using SD desk-microphone modelling.

Weighting Scheme		SI Head Models		
		uw	*cfw*	*cw*
Precision	5 docs	0.232	0.244	0.272
	10 docs	0.182	0.192	0.232
	15 docs	0.148	0.175	0.199
	20 docs	0.130	0.154	0.180
Av Precision		0.241	0.263	0.300

Table 10: VMR1a retrieval precision for WS using SI head-microphone modelling.

Weighting Scheme		SI Head Models		
		uw	*cfw*	*cw*
Precision	5 docs	0.283	0.321	0.363
	10 docs	0.240	0.283	0.313
	15 docs	0.221	0.258	0.268
	20 docs	0.204	0.231	0.245
Av Precision		0.251	0.291	0.301

Table 11: VMR1b retrieval precision for WS using SI head-microphone modelling.

Weighting Scheme		SI Desk Models		
		uw	*cfw*	*cw*
Precision	5 docs	0.192	0.228	0.256
	10 docs	0.152	0.178	0.212
	15 docs	0.128	0.153	0.172
	20 docs	0.113	0.130	0.148
Av Precision		0.184	0.219	0.275

Table 12: VMR1a retrieval precision for WS using SI desk-microphone modelling.

Weighting Scheme		SI Desk Models		
		uw	*cfw*	*cw*
Precision	5 docs	0.250	0.271	0.300
	10 docs	0.227	0.250	0.271
	15 docs	0.193	0.219	0.246
	20 docs	0.173	0.196	0.232
Av Precision		0.214	0.249	0.267

Table 13: VMR1b retrieval precision for WS using SI desk-microphone modelling.

		SI + R13 Head Models		
Weighting Scheme		*uw*	*cfw*	*cw*
Precision	5 docs	0.212	0.224	0.272
	10 docs	0.180	0.208	0.236
	15 docs	0.159	0.185	0.208
	20 docs	0.142	0.163	0.185
Av Precision		0.234	0.274	0.324

Table 14: VMR1a retrieval precision for WS using SI head-microphone modelling adapted with R13 head-microphone data.

		SI + R13 Head Models		
Weighting Scheme		*uw*	*cfw*	*cw*
Precision	5 docs	0.288	0.325	0.321
	10 docs	0.240	0.273	0.296
	15 docs	0.226	0.249	0.271
	20 docs	0.207	0.244	0.254
Av Precision		0.245	0.283	0.305

Table 15: VMR1b retrieval precision for WS using SI head-microphone modelling adapted with R13 head-microphone data.

		SI + R75 Head Models		
Weighting Scheme		*uw*	*cfw*	*cw*
Precision	5 docs	0.244	0.264	0.300
	10 docs	0.204	0.218	0.246
	15 docs	0.173	0.199	0.219
	20 docs	0.154	0.171	0.188
Av Precision		0.256	0.294	0.338

Table 16: VMR1a retrieval precision for WS using SI head-microphone modelling adapted with R75 head-microphone data.

		SI + R75 Head Models		
Weighting Scheme		*uw*	*cfw*	*cw*
Precision	5 docs	0.300	0.329	0.358
	10 docs	0.258	0.296	0.323
	15 docs	0.236	0.278	0.288
	20 docs	0.216	0.256	0.260
Av Precision		0.270	0.308	0.334

Table 17: VMR1b retrieval precision for WS using SI head-microphone modelling adapted with R75 head-microphone data.

Weighting Scheme		SI + R13 Desk Models		
		uw	*cfw*	*cw*
Precision	5 docs	0.172	0.196	0.240
	10 docs	0.152	0.156	0.184
	15 docs	0.127	0.137	0.164
	20 docs	0.106	0.122	0.141
Av Precision		0.180	0.213	0.243

Table 18: VMR1a retrieval precision for WS using SI desk-microphone modelling adapted with R13 desk-microphone data.

Weighting Scheme		SI + R13 Desk Models		
		uw	*cfw*	*cw*
Precision	5 docs	0.234	0.271	0.304
	10 docs	0.213	0.231	0.267
	15 docs	0.179	0.208	0.243
	20 docs	0.166	0.202	0.227
Av Precision		0.203	0.251	0.279

Table 19: VMR1b retrieval precision for WS using SI desk-microphone modelling adapted with R13 desk-microphone data.

Weighting Scheme		SI + R75 Desk Models		
		uw	*cfw*	*cw*
Precision	5 docs	0.196	0.224	0.244
	10 docs	0.142	0.186	0.212
	15 docs	0.133	0.157	0.176
	20 docs	0.120	0.139	0.152
Av Precision		0.192	0.232	0.266

Table 20: VMR1a retrieval precision for WS using SI desk-microphone modelling adapted with R75 desk-microphone data.

Weighting Scheme		SI + R75 Desk Models		
		uw	*cfw*	*cw*
Precision	5 docs	0.275	0.313	0.346
	10 docs	0.244	0.271	0.294
	15 docs	0.214	0.244	0.278
	20 docs	0.189	0.217	0.242
Av Precision		0.234	0.276	0.295

Table 21: VMR1b retrieval precision for WS using SI desk-microphone modelling adapted with R75 desk-microphone data.

Weighting Scheme			Average Precision		
			uw	*cfw*	*cw*
Text	Avg. Prec.		0.293	0.332	0.358
	(relative)		100%	100%	100%
Spoken Documents	SD	Head	88.4%	88.9%	88.3%
		Desk	82.3%	85.2%	83.5%
	SI	Head	82.3%	79.2%	83.8%
		Desk	62.8%	66.0%	76.8%
	SI + R13	Head	79.9%	82.5%	90.5%
		Desk	61.4%	64.2%	67.9%
	SI + R75	Head	87.4%	88.6%	94.4%
		Desk	65.5%	69.9%	74.3%

Table 22: Summary of VMR1a retrieval average precision for WS.

Weighting Scheme			Average Precision		
			uw	*cfw*	*cw*
Text	Avg. Prec.		0.296	0.332	0.346
	(relative)		100%	100%	100%
Spoken Documents	SD	Head	89.5%	94.0%	95.4%
		Desk	85.8%	89.2%	91.0%
	SI	Head	84.8%	87.7%	87.0%
		Desk	72.3%	75.0%	77.2%
	SI + R13	Head	82.8%	85.2%	88.2%
		Desk	68.6%	75.6%	80.6%
	SI + R75	Head	91.2%	92.8%	96.5%
		Desk	79.1%	83.1%	85.3%

Table 23: Summary of VMR1b retrieval average precision for WS.

well correlated to WS behaviour and that R75 data gives retrieval performance equivalent to SD models. This latter result is not surprising in view of the amount of adaptation data, which is nearly all the training data used for the initial SD models. Although not shown here, improvement is most dramatic for our north American speaker where the poorly matching British English SI models have been adapted to his north American speaking style. These conclusions are reinforced by the corresponding results for desk microphone data shown in Tables 18, 19, 20 and 21 which follow the same trends, albeit with overall lower retrieval performance for the reasons noted for SD retrieval.

5.1.5 Concluding Observations on WS

Tables 22 and 23 show a summary of average precision for our WS systems. Overall retrieval using fixed vocabulary keyword spotting has been shown to be reasonably effective. Retrieval performance is clearly well correlated to acoustic WS performance, with obvious implications for the design of word spotters for retrieval applications. As stated

Weighting Scheme		Text		
		uw	*cfw*	*cw*
Precision	5 docs	0.584	0.644	0.692
	10 docs	0.406	0.440	0.442
	15 docs	0.303	0.319	0.332
	20 docs	0.241	0.250	0.264
Av Precision		0.600	0.671	0.718

Table 24: VMR1a retrieval precision for open-vocabulary manual text transcription.

Weighting Scheme		Text		
		uw	*cfw*	*cw*
Precision	5 docs	0.392	0.375	0.371
	10 docs	0.313	0.308	0.344
	15 docs	0.279	0.292	0.308
	20 docs	0.250	0.271	0.290
Av Precision		0.327	0.352	0.368

Table 25: VMR1b retrieval precision for open-vocabulary manual text transcription.

previously the WS approach is restricted by the need to know the required search vocabulary in advance of recognition (and, of course, retrieval). Overall speaker adaptation is less effective in improving retrieval performance for the SI desk-microphone models. This is probably a reflection of similar behaviour for the WS FOM noted in Section 3.1.

5.2 Open Vocabulary Retrieval : PLS Tests

Work described in this section and the next is taken from stage 3 of the VMR project and covers investigations into open vocabulary retrieval. Since this work has not previously been reported in depth, the results given here are rather more comprehensive than those for fixed vocabulary retrieval. In this section we give first the reference results for text retrieval, and then results for phone lattice spotting. In the next section, 5.3, we consider open vocabulary retrieval using large vocabulary recognition.

5.2.1 Text Retrieval

Once again a reference standard is defined by computing retrieval performance for text transcriptions of the documents. Tables 24 and 25 show open-vocabulary text retrieval performance for VMR1a and VMR1b respectively. These results were obtained using the standard van Rijsbergen stop list [van Rijsbergen, 1979] and, of course, Porter suffix stripping. It can be observed that retrieval performance for VMR1a is significantly improved by the open-vocabulary compared to the fixed WS vocabulary shown in Table 4; although perhaps surprisingly, given the increased average number of terms in each query, the performance here for VMR1b is not much better than that shown in Table 5. However it should be remembered that the fixed keywords chosen were well matched to the VMR1 domain and many of the additional terms may be of only limited utility for these documents. Also the proportional increase in the query length is much higher for VMR1a. The open text VMR1a queries increase on average from 5.7 terms to 19.0 terms, whereas VMR1b increase from 2.6 to only 7.4 terms.

Evaluating Different Stop Word Lists A number of stop word lists were investigated as alternatives to the standard van Rijsbergen list. Although still widely used this list is now quite old and it appears that with increased availability of computer storage space many contemporary systems use much smaller lists. Four lists supplied to us by City University were investigated in this study. These were lists found to be generally useful by the team at City and were in no way tuned to the VMR1 domain. The number of stop words in each list is as follows:

	List 1	List 2	List 3	List 4	van R
No of Stop Words	229	32	7	3	290

Tables 26, 27, 28, 29, 30, 31, 32 and 33 show retrieval performance for the four alternative stop word lists. From these results it would appear that there are no significant variations in retrieval performance for VMR1a and VMR1b arising from using the different stop word lists. Of course, the size of the VMR1 archive means that this result itself cannot be taken as significant. Bearing this important point in mind some observations on the results observed can be made as follows. Overall the trend would appear to be that the best retrieval performance is achieved using the original van Rijsbergen list, and that performance is on average progressively degraded as the stop word list is shortened. This effect is most clearly seen to the *uw* scheme, is much reduced for *cfw* ; while variation for *cw* could be attributed to noise.

Overall this result is slightly surprising given the apparent current popularity of short stop word lists. However, one particular features of these shorter lists is that they enable more sophisticated retrieval strategies to be examined, e.g. ones using phrases, and it may be that we would see a greater effect if we were to investigate such techniques. An alternative reasonable explanation for these results arises from the term weighting methods, and in particular the collection frequency weighting (*cfw*) component. The *cfw* for term i is calculated as $cfw(i) = \log N - \log n$ where is N is the number of documents in the collection and n is the number of documents in which i is present. Where n is relatively large, as will often be the case for potential stop words, there will be little change in the size of $cfw(i)$ as the size of the document archive increases. However, for rarer content bearing words the relative difference between N and n is likely to increase rapidly when the size of N increases. Thus for large archives, such as those currently used in most text retrieval experimentation, the relative impact of relatively common search terms on the overall matching score between a document and a query will frequently be much reduced.

5.2.2 Phone Lattice Spotting

The following sections give results for PLS with different acoustic models. The first three experiments show performance for the head microphone data and the last experiment for desk microphone data.

		Text		
Weighting Scheme		*uw*	*cfw*	*cw*
Precision	5 docs	0.576	0.628	0.684
	10 docs	0.392	0.430	0.440
	15 docs	0.295	0.311	0.327
	20 docs	0.236	0.247	0.262
Av Precision		0.602	0.656	0.705

Table 26: VMR1a retrieval precision for open-vocabulary manual text transcription with stop word list 1.

		Text		
Weighting Scheme		*uw*	*cfw*	*cw*
Precision	5 docs	0.379	0.371	0.375
	10 docs	0.331	0.315	0.333
	15 docs	0.275	0.293	0.317
	20 docs	0.255	0.279	0.288
Av Precision		0.321	0.353	0.363

Table 27: VMR1b retrieval precision for open-vocabulary manual text transcription with stop word list 1.

		Text		
Weighting Scheme		*uw*	*cfw*	*cw*
Precision	5 docs	0.488	0.624	0.684
	10 docs	0.358	0.428	0.452
	15 docs	0.269	0.312	0.332
	20 docs	0.215	0.245	0.264
Av Precision		0.512	0.654	0.720

Table 28: VMR1a retrieval precision for open-vocabulary manual text transcription with stop word list 2.

		Text		
Weighting Scheme		*uw*	*cfw*	*cw*
Precision	5 docs	0.321	0.350	0.396
	10 docs	0.275	0.306	0.338
	15 docs	0.244	0.282	0.307
	20 docs	0.225	0.253	0.284
Av Precision		0.264	0.324	0.357

Table 29: VMR1b retrieval precision for open-vocabulary manual text transcription with stop word list 2.

		Text		
Weighting Scheme		*uw*	*cfw*	*cw*
Precision	5 docs	0.476	0.624	0.700
	10 docs	0.326	0.428	0.452
	15 docs	0.249	0.307	0.336
	20 docs	0.210	0.245	0.262
Av Precision		0.493	0.655	0.721

Table 30: VMR1a retrieval precision for open-vocabulary manual text transcription with stop word list 3.

		Text		
Weighting Scheme		*uw*	*cfw*	*cw*
Precision	5 docs	0.325	0.354	0.408
	10 docs	0.279	0.308	0.342
	15 docs	0.240	0.282	0.313
	20 docs	0.216	0.251	0.279
Av Precision		0.257	0.325	0.354

Table 31: VMR1b retrieval precision for open-vocabulary manual text transcription with stop word list 3.

		Text		
Weighting Scheme		*uw*	*cfw*	*cw*
Precision	5 docs	0.460	0.624	0.700
	10 docs	0.318	0.428	0.454
	15 docs	0.244	0.307	0.332
	20 docs	0.202	0.245	0.261
Av Precision		0.476	0.655	0.719

Table 32: VMR1a retrieval precision for open-vocabulary manual text transcription with stop word list 4.

		Text		
Weighting Scheme		*uw*	*cfw*	*cw*
Precision	5 docs	0.328	0.354	0.404
	10 docs	0.279	0.310	0.340
	15 docs	0.238	0.285	0.315
	20 docs	0.212	0.251	0.278
Av Precision		0.253	0.325	0.358

Table 33: VMR1b retrieval precision for open-vocabulary manual text transcription with stop word list 4.

Weighting Scheme		SD Head Models				
		uw	*cfw*	*cw*		
				Terms	Time	Phones
Precision	5 docs	0.366	0.408	0.392	0.480	0.472
	10 docs	0.252	0.312	0.286	0.342	0.340
	15 docs	0.201	0.235	0.235	0.255	0.259
	20 docs	0.167	0.195	0.193	0.211	0.211
Av Precision		0.366	0.427	0.398	0.490	0.495

Table 34: VMR1a retrieval precision for SD head-microphone monophone models PLS.

Weighting Scheme		SD Head Models				
		uw	*cfw*	*cw*		
				Terms	Time	Phones
Precision	5 docs	0.329	0.288	0.317	0.342	0.338
	10 docs	0.267	0.254	0.240	0.279	0.288
	15 docs	0.213	0.225	0.214	0.243	0.246
	20 docs	0.197	0.212	0.198	0.225	0.222
Av Precision		0.262	0.285	0.284	0.311	0.315

Table 35: VMR1b retrieval precision for SD head-microphone monophone models PLS.

Weighting Scheme		SI Head Models				
		uw	*cfw*	*cw*		
				Words	Time	Phones
Precision	5 docs	0.252	0.304	0.324	0.376	0.372
	10 docs	0.186	0.222	0.234	0.254	0.268
	15 docs	0.152	0.185	0.184	0.208	0.212
	20 docs	0.125	0.149	0.151	0.175	0.177
Av Precision		0.253	0.318	0.320	0.386	0.390

Table 36: VMR1a retrieval precision for SI head-microphone monophone models PLS.

Weighting Scheme		SI Head Models				
		uw	*cfw*	*cw*		
				Words	Time	Phones
Precision	5 docs	0.200	0.233	0.238	0.250	0.242
	10 docs	0.160	0.183	0.190	0.190	0.200
	15 docs	0.146	0.168	0.175	0.199	0.200
	20 docs	0.148	0.157	0.170	0.183	0.185
Av Precision		0.174	0.199	0.208	0.216	0.222

Table 37: VMR1b retrieval precision for SI head-microphone monophone models PLS.

Speaker-Dependent Monophone Modelling Tables 34 and 35 show PLS retrieval performance for SD head-microphone monophone modelling, including various different ways of defining document length. It is shown clearly that using only the terms present in the query to represent the document length is ineffective and that the count of phones in the most likely path from the Viterbi decoding is the best measure of document length available. This gives an improvement of more than 20% for VMR1a and around 10% for VMR1b.

Speaker-Independent Monophone Modelling Tables 36 and 37 show retrieval performance for PLS with SI monophones. Performance is considerably reduced relative to PLS with SD modelling, but this is to be expected due to the degraded PLS word spotting accuracy noted in Section 3.2. Similar percentage improvements in retrieval performance can be observed as with the SD monophones when the document length is measured in phones or time.

Speaker-Independent Biphone Modelling Tables 38 and 39 show that the modelling of context provided by the biphone models produces an anticipated improvement in retrieval performance. Although the retrieval performance is not as good as for SD monophone modelling, the biphones models represent the best SI system available. It would be expected that SD biphones would give improved retrieval performance over any of the PLS models considered here, unfortunately insufficient speaker-dependent acoustic training data meant that this experiment could not be carried out.

Tables 40 and 41 show PLS retrieval performance for desk-microphone data SI biphones. These are the models used in the final VMR demonstration system. As for WS, desk-microphone models for PLS perform significantly worse than the head-microphone system. Based on the head-microphone results, we would anticipate that for the desk-microphone data SD monophones and SI monophones would perform better and worse respectively to the biphones, however this hypothesis was not tested.

Additionally, SI PLS performance could probably be improved by the application of speaker adaptation, followed by a rescoring of the lattice or perhaps a complete reprocessing of the data.

5.2.3 Concluding Observations on PLS

Tables 42 and 43 show a summary of retrieval performance using PLS for VMR1a and VMR1b respectively. From these comparative results it is clear that there is once again a strong correlation between speech recognition modelling quality and retrieval performance. It is shown that improvements in acoustic modelling can produce significant improvements in retrieval performance and thus continued research into acoustic modelling is clearly important.

SI modelling is clearly preferable in practice, but SD results are also given to illustrate the increased potential of PLS where superior acoustic modelling is available. Overall, PLS has been shown to be a useful technique in retrieval of spoken documents. However, it is clear that much additional research into PLS methods remains to be carried out. It should be noted that in these experiments search words were only searched for in the form originally specified in the request. As was described in detail in Section 4.2 there

Weighting Scheme		SI Head Models				
		uw	*cfw*	*cw*		
				Words	Time	Phones
Precision	5 docs	0.308	0.400	0.396	0.464	0.468
	10 docs	0.226	0.286	0.276	0.314	0.322
	15 docs	0.185	0.224	0.220	0.248	0.247
	20 docs	0.154	0.182	0.183	0.204	0.204
Av Precision		0.319	0.411	0.400	0.462	0.470

Table 38: VMR1a retrieval precision for SI head-microphone biphone models PLS.

Weighting Scheme		SI Head Models				
		uw	*cfw*	*cw*		
				Words	Time	Phones
Precision	5 docs	0.296	0.279	0.233	0.313	0.317
	10 docs	0.235	0.254	0.223	0.248	0.254
	15 docs	0.199	0.226	0.206	0.226	0.236
	20 docs	0.171	0.198	0.200	0.206	0.205
Av Precision		0.224	0.262	0.248	0.269	0.277

Table 39: VMR1b retrieval precision for SI head-microphone biphone models PLS.

Weighting Scheme		SI Desk Models				
		uw	*cfw*	*cw*		
				Words	Time	Phones
Precision	5 docs	0.272	0.356	0.316	0.364	0.360
	10 docs	0.208	0.246	0.220	0.260	0.266
	15 docs	0.168	0.187	0.181	0.204	0.204
	20 docs	0.143	0.158	0.153	0.172	0.176
Av Precision		0.286	0.362	0.321	0.374	0.377

Table 40: VMR1a retrieval precision for SI desk-microphone biphone models PLS.

Weighting Scheme		SI Desk Models				
		uw	*cfw*	*cw*		
				Words	Time	Phones
Precision	5 docs	0.242	0.217	0.200	0.229	0.225
	10 docs	0.210	0.215	0.181	0.231	0.235
	15 docs	0.197	0.196	0.172	0.210	0.207
	20 docs	0.181	0.185	0.165	0.188	0.188
Av Precision		0.183	0.206	0.201	0.216	0.226

Table 41: VMR1b retrieval precision for SI desk-microphone biphone models PLS.

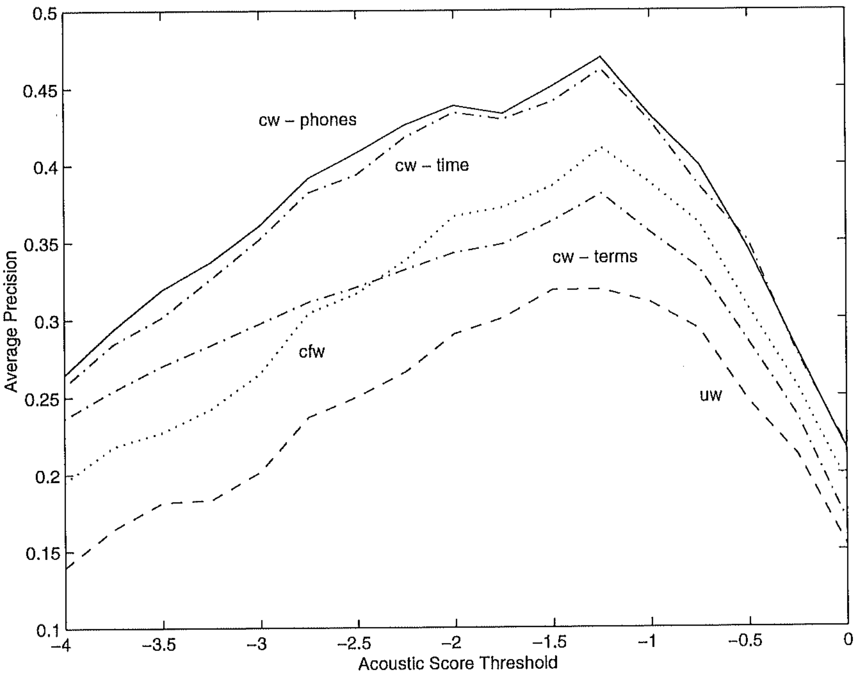

Figure 1: Retrieval performance against acoustic threshold for various term weighting schemes with head-microphone SI biphones for VMR1a.

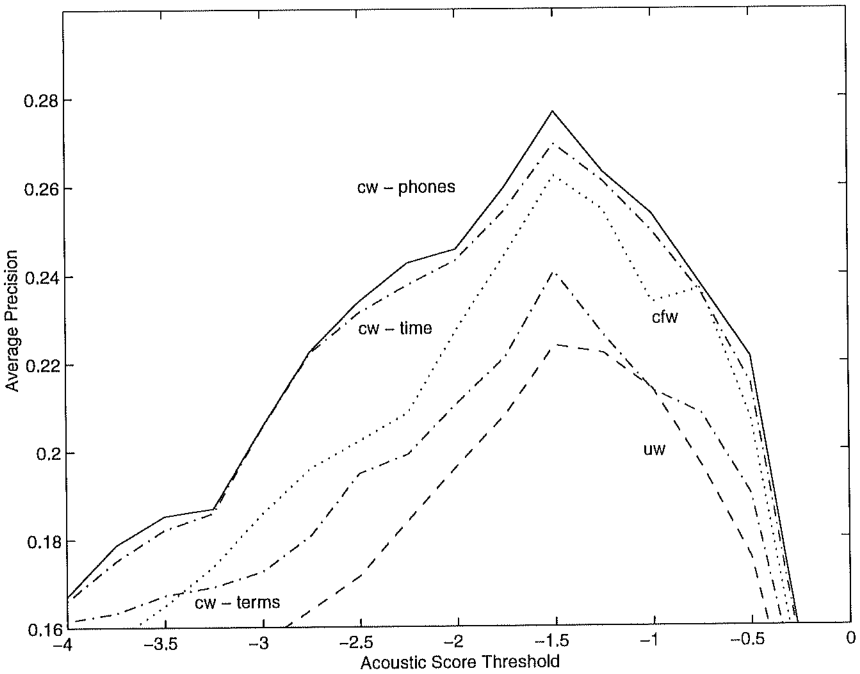

Figure 2: Retrieval performance against acoustic threshold for various term weighting schemes with head-microphone SI biphones for VMR1b.

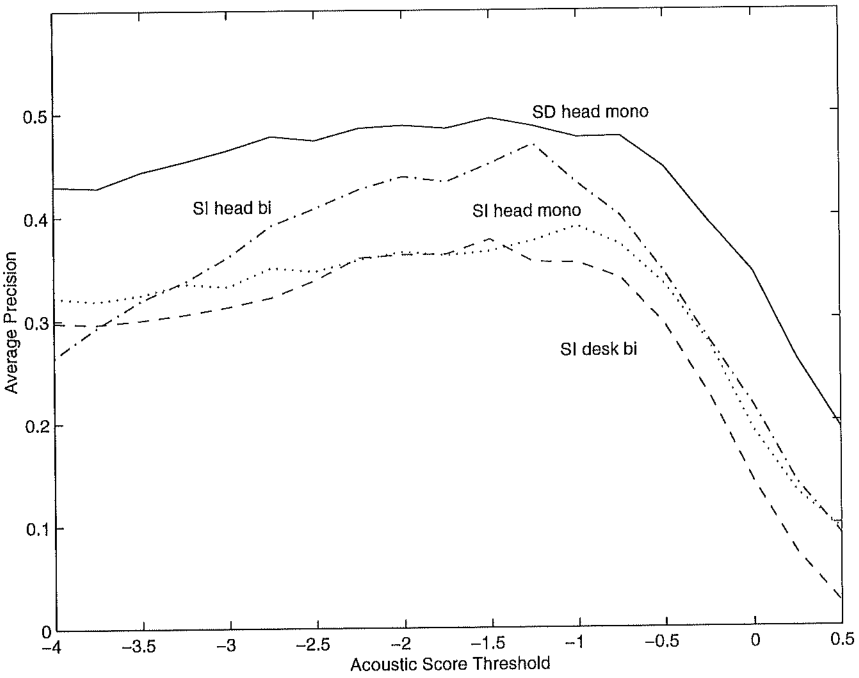

Figure 3: PLS retrieval performance against acoustic score with *cw* weighting for different acoustic models for VMR1a.

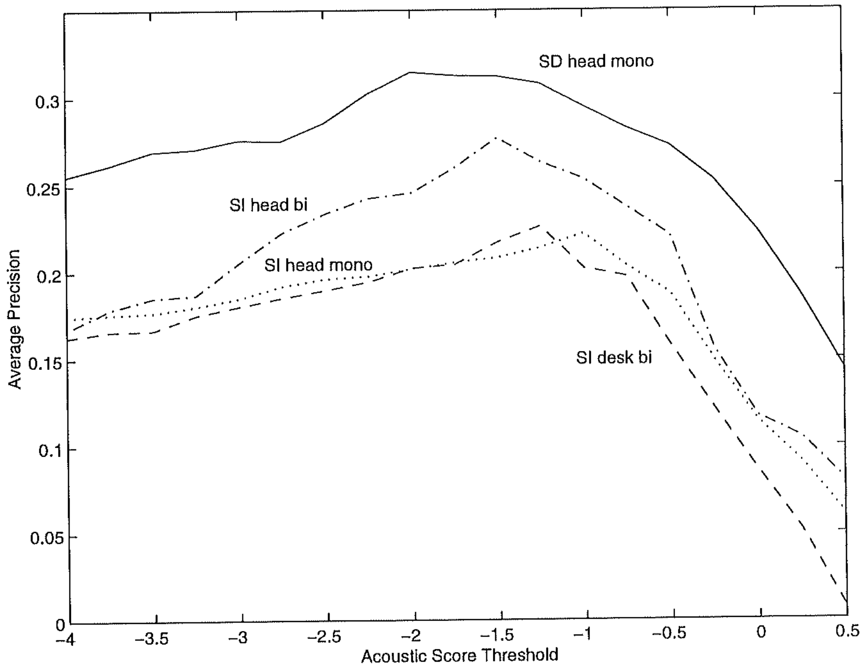

Figure 4: PLS retrieval performance against acoustic score with *cw* weighting for different acoustic models for VMR1b.

Weighting Scheme			Average Precision		
			uw	*cfw*	*cw*
Text	Avg. Prec.		0.600	0.671	0.718
	(relative)		100%	100%	100%
Spoken Documents	SD	Head Monophones	61.0%	63.6%	68.9%
	SI	Head Monophones	42.2%	47.4%	54.3%
		Head Biphones	53.2%	61.3%	65.5%
		Desk Biphones	47.7%	53.9%	52.5%

Table 42: Summary of VMR1a retrieval average precision for PLS.

Weighting Scheme			Average Precision		
			uw	*cfw*	*cw*
Text	Avg. Prec.		0.327	0.352	0.368
	(relative)		100%	100%	100%
Spoken Documents	SD	Head Monophones	80.1%	81.0%	85.6%
	SI	Head Monophones	53.2%	56.5%	60.3%
		Head Biphones	68.5%	74.4%	75.3%
		Desk Biphones	56.0%	58.5%	61.4%

Table 43: Summary of VMR1b retrieval average precision for PLS.

are several other approaches possible for this, but shortness of time prevented additional investigation.

Figures 1 and 2 show average retrieval precision performance against acoustic threshold for VMR1a and VMR1b respectively with SI biphone models. It can be seen in both cases that *cw* weighting with document length measured in time or phones gives better overall performance.

Figures 3 and 4 show average precision retrieval performance against score threshold for VMR1a and VMR1b respectively for the different acoustic models. These figures compare retrieval performance using *cw* weighting using document length measured in phones for SD monophones, SI monophones and SI biphones. The figures show that not only does the SD model give the best single threshold average precision in each case, but also that it is in general more robust to the exact choice of threshold.

A very positive observation is that retrieval performance for desk-microphone biphones, i.e. final VMR demonstration system is around 60% of text retrieval performance when using *cw* weighting.

All these figures indicate that the general quality of retrieval is improved when better acoustic models are used.

5.3 Open Vocabulary Retrieval : LVR Tests

We now consider retrieval performance for the alternative model of word recognition, using LVR.

Weighting Scheme		Text		
		uw	*cfw*	*cw*
Precision	5 docs	0.560	0.632	0.684
	10 docs	0.388	0.426	0.434
	15 docs	0.289	0.315	0.327
	20 docs	0.234	0.247	0.261
Av Precision		0.576	0.653	0.703

Table 44: VMR1a retrieval precision for 20K WSJ vocabulary manual text transcription.

Weighting Scheme		Text		
		uw	*cfw*	*cw*
Precision	5 docs	0.346	0.313	0.321
	10 docs	0.281	0.277	0.294
	15 docs	0.257	0.257	0.272
	20 docs	0.227	0.242	0.258
Av Precision		0.299	0.312	0.325

Table 45: VMR1b retrieval precision for 20K WSJ vocabulary manual text transcription.

Experiments beyond this point are concerned only with the head-microphone data since LVR experiments using desk-microphone data were not carried out.

5.3.1 Large Vocabulary Speech Recognition

This section gives retrieval results using the 20K WSJ LVR system. Since this recogniser has an OOV rate in excess of 3% for the VMR1 corpus we anticipate that retrieval performance may be somewhat degraded merely by the absence of some search terms. Hence the first experiment shows manual text transcription retrieval performance where the vocabulary has been limited to that of the 20K recogniser. The second experiment shows retrieval performance using the transcription output from the 20K WSJ LVR system.

20K Text Retrieval Tables 44 and 45 show performance for the *manual* transcriptions for VMR1a and VMR1b respectively, when using the WSJ 20K vocabulary and the van Rijsbergen stop list. As anticipated retrieval performance is reduced in all cases relative to fully open-vocabulary text retrieval as shown in Tables 24 and 25.

Tables 46, 47, 48, 49, 50, 51, 52 and 53 show WSJ 20K text retrieval performance for the four stop word lists examined for open text retrieval. Similar behaviour with respect to the alternative stop word lists is observed for 20K text as that noted for open-vocabulary text. One slight difference is an apparent slight improvement in VMR1b retrieval performance for *cw* weighting when using the shorter stop word lists. This increase is small and cannot be regarded as significant from these results. A possible explanation for this effect is as follows. Having removed some important search terms by restricting the search vocabulary to the WSJ 20K, it may be that additional terms not removed in the shorted lists may act

		Text		
Weighting Scheme		*uw*	*cfw*	*cw*
Precision	5 docs	0.572	0.620	0.676
	10 docs	0.390	0.412	0.430
	15 docs	0.287	0.305	0.327
	20 docs	0.231	0.244	0.258
Av Precision		0.593	0.646	0.691

Table 46: VMR1a retrieval precision for manual text transcription reduced to 20K WSJ vocabulary with stop word list 1.

		Text		
Weighting Scheme		*uw*	*cfw*	*cw*
Precision	5 docs	0.367	0.350	0.367
	10 docs	0.313	0.300	0.329
	15 docs	0.267	0.281	0.313
	20 docs	0.246	0.270	0.282
Av Precision		0.306	0.338	0.354

Table 47: VMR1b retrieval precision for manual text transcription reduced to 20K WSJ vocabulary with stop word list 1.

		Text		
Weighting Scheme		*uw*	*cfw*	*cw*
Precision	5 docs	0.488	0.612	0.668
	10 docs	0.350	0.412	0.442
	15 docs	0.259	0.309	0.331
	20 docs	0.212	0.242	0.261
Av Precision		0.496	0.639	0.707

Table 48: VMR1a retrieval precision for manual text transcription reduced to 20K WSJ vocabulary with stop word list 2.

		Text		
Weighting Scheme		*uw*	*cfw*	*cw*
Precision	5 docs	0.313	0.333	0.383
	10 docs	0.263	0.292	0.327
	15 docs	0.233	0.265	0.297
	20 docs	0.216	0.243	0.277
Av Precision		0.256	0.308	0.342

Table 49: VMR1b retrieval precision for manual text transcription reduced to 20K WSJ vocabulary with stop word list 2.

Weighting Scheme		Text		
		uw	*cfw*	*cw*
Precision	5 docs	0.476	0.616	0.684
	10 docs	0.316	0.412	0.444
	15 docs	0.244	0.303	0.333
	20 docs	0.203	0.240	0.261
Av Precision		0.477	0.640	0.707

Table 50: VMR1a retrieval precision for manual text transcription reduced to 20K WSJ vocabulary with stop word list 3.

Weighting Scheme		Text		
		uw	*cfw*	*cw*
Precision	5 docs	0.308	0.338	0.400
	10 docs	0.267	0.294	0.331
	15 docs	0.231	0.267	0.310
	20 docs	0.207	0.241	0.270
Av Precision		0.249	0.309	0.339

Table 51: VMR1b retrieval precision for manual text transcription reduced to 20K WSJ vocabulary with stop word list 3.

Weighting Scheme		Text		
		uw	*cfw*	*cw*
Precision	5 docs	0.464	0.616	0.684
	10 docs	0.312	0.412	0.444
	15 docs	0.237	0.303	0.329
	20 docs	0.195	0.240	0.259
Av Precision		0.461	0.640	0.704

Table 52: VMR1a retrieval precision for manual text transcription reduced to 20K WSJ vocabulary with stop word list 4.

Weighting Scheme		Text		
		uw	*cfw*	*cw*
Precision	5 docs	0.313	0.338	0.392
	10 docs	0.267	0.296	0.329
	15 docs	0.228	0.269	0.307
	20 docs	0.203	0.241	0.269
Av Precision		0.245	0.310	0.341

Table 53: VMR1b retrieval precision for manual text transcription reduced to 20K WSJ vocabulary with stop word list 4.

Weighting Scheme		20K WSJ LVR		
		uw	*cfw*	*cw*
Precision	5 docs	0.472	0.540	0.568
	10 docs	0.316	0.340	0.370
	15 docs	0.235	0.255	0.273
	20 docs	0.190	0.211	0.219
Av Precision		0.475	0.523	0.574

Table 54: VMR1a retrieval precision for 20K WSJ LVR.

Weighting Scheme		20K WSJ LVR		
		uw	*cfw*	*cw*
Precision	5 docs	0.254	0.271	0.304
	10 docs	0.213	0.238	0.248
	15 docs	0.204	0.219	0.239
	20 docs	0.184	0.193	0.217
Av Precision		0.225	0.246	0.264

Table 55: VMR1b retrieval precision for 20K WSJ LVR.

as useful search terms when *cw* weighting is used. This hypothesis would obviously have to be tested on a much larger archive before it could be accepted. An alternative explanation is that the variation in *cw* results arises from slightly changes in the term weights arising from changes in measured document length when different numbers of terms are removed. More extensive experiments would be needed to investigate these possibilities, these were not attempted here.

Since overall the variation in retrieval results is minimal, results obtained with the van Rijsbergen list are used as the standard for comparison against later results.

20K Wall Street Journal recogniser As mentioned in Section 3.3, our tests with LVR used the standard WSJ 20K vocabulary and language model. As the output of the recognition is a transcription of the original speech, the actual search tests are done on the texts given by this *automated* transcription. Such texts may naturally be subjected to the use of stop lists.

Tables 54 and 55 show VMR1a and VMR1b retrieval performance for the automated transcription output of the 20K WSJ recogniser using the van Rijsbergen stop list. Results here are further degraded relative to open-vocabulary text by the speech recognition errors in the LVR system.

Tables 56, 57, 58, 59, 60, 61, 62 and 63 again show retrieval performance with the alternative stop word lists. Similar trends are observed for different stop word lists to those already noted for previous experiments with the different lists. Although overall it appears that the shorter stop word lists are less preferable for LVR than for the text systems. As noted before these results cannot be taken as significant, however a clear

		20K WSJ LVR		
Weighting Scheme		*uw*	*cfw*	*cw*
Precision	5 docs	0.460	0.520	0.568
	10 docs	0.304	0.338	0.356
	15 docs	0.236	0.247	0.271
	20 docs	0.195	0.203	0.217
Av Precision		0.470	0.515	0.566

Table 56: VMR1a retrieval precision for 20K WSJ LVR for stop word list 1.

		20K WSJ LVR		
Weighting Scheme		*uw*	*cfw*	*cw*
Precision	5 docs	0.225	0.258	0.296
	10 docs	0.219	0.231	0.256
	15 docs	0.206	0.210	0.233
	20 docs	0.189	0.195	0.219
Av Precision		0.225	0.238	0.261

Table 57: VMR1b retrieval precision for 20K WSJ LVR for stop word list 1.

		20K WSJ LVR		
Weighting Scheme		*uw*	*cfw*	*cw*
Precision	5 docs	0.380	0.496	0.580
	10 docs	0.242	0.330	0.372
	15 docs	0.192	0.252	0.281
	20 docs	0.156	0.204	0.225
Av Precision		0.347	0.492	0.583

Table 58: VMR1a retrieval precision for 20K WSJ LVR for stop word list 2.

		20K WSJ LVR		
Weighting Scheme		*uw*	*cfw*	*cw*
Precision	5 docs	0.217	0.217	0.275
	10 docs	0.188	0.213	0.250
	15 docs	0.167	0.206	0.219
	20 docs	0.151	0.191	0.207
Av Precision		0.182	0.217	0.243

Table 59: VMR1b retrieval precision for 20K WSJ LVR for stop word list 2.

		20K WSJ LVR		
Weighting Scheme		*uw*	*cfw*	*cw*
Precision	5 docs	0.360	0.504	0.580
	10 docs	0.236	0.330	0.378
	15 docs	0.183	0.251	0.288
	20 docs	0.146	0.201	0.225
Av Precision		0.336	0.494	0.581

Table 60: VMR1a retrieval precision for 20K WSJ LVR for stop word list 3.

		20K WSJ LVR		
Weighting Scheme		*uw*	*cfw*	*cw*
Precision	5 docs	0.221	0.225	0.283
	10 docs	0.194	0.213	0.242
	15 docs	0.167	0.208	0.221
	20 docs	0.146	0.190	0.204
Av Precision		0.175	0.216	0.243

Table 61: VMR1b retrieval precision for 20K WSJ LVR for stop word list 3.

		20K WSJ LVR		
Weighting Scheme		*uw*	*cfw*	*cw*
Precision	5 docs	0.356	0.504	0.584
	10 docs	0.236	0.328	0.380
	15 docs	0.180	0.251	0.285
	20 docs	0.144	0.201	0.225
Av Precision		0.325	0.494	0.580

Table 62: VMR1a retrieval precision for 20K WSJ LVR for stop word list 4.

		20K WSJ LVR		
Weighting Scheme		*uw*	*cfw*	*cw*
Precision	5 docs	0.233	0.238	0.292
	10 docs	0.190	0.213	0.250
	15 docs	0.165	0.211	0.218
	20 docs	0.144	0.191	0.204
Av Precision		0.179	0.222	0.244

Table 63: VMR1b retrieval precision for 20K WSJ LVR for stop word list 4.

Weighting Scheme			Average Precision		
			uw	*cfw*	*cw*
Text	Full Vocab.	Avg. Prec.	0.600	0.671	0.718
		(relative)	100%	100%	100%
	20K Vocab.		96.0%	97.3%	97.9%
Spoken Documents	20K WSJ LVR		79.2%	77.9%	79.9%

Table 64: Summary of VMR1a retrieval average precision for 20K WSJ LVR.

Weighting Scheme			Average Precision		
			uw	*cfw*	*cw*
Text	Full Vocab.	Avg. Prec.	0.327	0.352	0.368
		(relative)	100%	100%	100%
	20K Vocab.		91.4%	88.6%	88.3%
Spoken Documents	20K WSJ LVR		68.8%	69.9%	71.7%

Table 65: Summary of VMR1b retrieval average precision for 20K WSJ LVR.

trend does seem to be apparent. This effect is probably attributable to recognition errors, particularly of shorter function words which are always harder to recognise accurately. More of these words are removed as the length of the stop word list is increased leading to an improvement in retrieval performance since the terms which remain are more likely to be correct. Since these variations are again small, the van Rijsbergen list is used in all following experiments.

5.3.2 Concluding Observations on LVR

Tables 64 and 65 show a summary of LVR retrieval performance for VMR1a and VMR1b respectively. For both VMR1a and VMR1b retrieval performance is degraded relative to open-text retrieval when the vocabulary is restricted to 20K text. This effect is noticibly smaller for VMR1a. The probable explanation for this is the increased redundancy associated with the much longer average request length in VMR1a. Hence queries derived from these requests will be less sensitive to the absence of individual search terms than the shorter VMR1b queries.

Retrieval performance is further degraded for the automated 20K LVR transcription. Again this degradation is much larger for VMR1b. But overall it is encouraging to note that for this 20K LVR system, which is not well matched to the VMR domain, a retrieval performance of 70% compared to text is obtained for VMR1b and 80% compared to text for VMR1a.

5.4 Combination Methods

The retrieval results reported in previous sections using different speech recognition systems each have advantages and disadvantages. The following experiments report results using the combination methods described in Section 4.3.

5.4.1 WS + 20K LVR

Experiments in this section concern the combination of fixed vocabulary WS and the 20K WSJ LVR system. This strategy of course assumes that there is an appropriate WS search vocabulary for a given application. It might for example be useful where LVR was used to generate a general transcription and WS was used to seek some specialised vocabulary, possibly at a later time, in advance of retrieval.

As noted in Section 4.3.2, the merged document length is calculated as the sum of all terms from both LVR and WS.

Data Fusion In data fusion the various scores from different types of index are combined for each document. Tables 66 and 67 show retrieval performance results from data fusion experiments combining 20K LVR with SD WS. Results are shown both for simple fusion where document matching scores are added together and normalised fusion where scores from each source are normalised relative to the highest score from that source before addition. Thus the effect of fusion is to derive a single output list from separate lists for each index, giving the results shown in the tables. (Note WS data used for all Data Fusion experiments reported here were in each case those at the acoustic score threshold which gave optimal retrieval performance for WS on its own.)

Tables 68 and 69 show retrieval performance for fusion using SI WS outputs and, as before, SI 20K WSJ LVR output. As for SI WS on its own, fusion retrieval performance is degraded in all cases for SI models relative to SD models.

The next experiments investigate the fusion retrieval performance using 20K WSJ output with adapted SI WS models. Tables 70 and 71 show fusion retrieval performance for WS models adapted using R13 data, and Tables 72 and 73 show fusion retrieval performance for WS models adapted using R75 data. (Note in each case for adapted WS the acoustic threshold and word insertion penalty were selected which gave the best available WS only retrieval performance.)

Observations Comparing these results to those for WS and LVR indexing systems on their own, overall it can be seen that the fusion of SD WS and LVR gives little improvement relative to LVR on its own for VMR1a and for fusion of SI WS with LVR there is actually a slight decrease. Conversely it can be seen that there is a good improvement in retrieval performance in all cases for VMR1b. As for the individual WS evaluation there is a clear correlation between acoustic WS FOM and retrieval performance. Retrieval performance is best for combination involving SD WS and least effective using SI WS. The adapted SI WS performs better as more adaptation data is used.

It can be seen that retrieval performance is better for VMR1a by direct score fusion and for VMR1b with the normalised score addition. This is probably again due to the effectiveness of LVR on its own for VMR1a compared to WS. In contrast for VMR1b WS and LVR are individually roughly comparable in retrieval performance, and thus it is reasonable that when their outputs are fused they should have equal weight to give the best performance. In the case of VMR1a the introduction of WS adversely affects the LVR scores and thus performs best where the WS component has the smaller influence, i.e. with direct score combination. Particularly for VMR1a, LVR match scores will on average have larger values than WS match scores since many more terms can be matched, so in

		Simp. Fuse			Norm. Fuse		
Weight Scheme		*uw*	*cfw*	*cw*	*uw*	*cfw*	*cw*
Prec.	5 docs	0.432	0.536	0.596	0.416	0.496	0.524
	10 docs	0.320	0.354	0.370	0.292	0.332	0.340
	15 docs	0.241	0.267	0.287	0.236	0.260	0.267
	20 docs	0.204	0.221	0.230	0.200	0.214	0.224
Av Precision		0.473	0.544	0.588	0.449	0.505	0.540

Table 66: VMR1a retrieval precision for data fusion combining 20K LVR and SD WS.

		Simp. Fuse			Norm. Fuse		
Weight Scheme		*uw*	*cfw*	*cw*	*uw*	*cfw*	*cw*
Prec.	5 docs	0.338	0.350	0.383	0.333	0.367	0.396
	10 docs	0.292	0.319	0.329	0.292	0.331	0.338
	15 docs	0.256	0.288	0.292	0.258	0.293	0.299
	20 docs	0.223	0.252	0.265	0.232	0.265	0.277
Av Precision		0.292	0.318	0.347	0.295	0.332	0.352

Table 67: VMR1b retrieval precision for data fusion combining 20K LVR and SD WS.

		Simp. Fuse			Norm. Fuse		
Weight Scheme		*uw*	*cfw*	*cw*	*uw*	*cfw*	*cw*
Prec.	5 docs	0.452	0.548	0.588	0.412	0.472	0.524
	10 docs	0.318	0.352	0.374	0.284	0.320	0.342
	15 docs	0.244	0.263	0.280	0.232	0.240	0.267
	20 docs	0.195	0.210	0.225	0.193	0.201	0.215
Av Precision		0.468	0.538	0.591	0.426	0.482	0.521

Table 68: VMR1a retrieval precision for data fusion combining 20K LVR and SI WS.

		Simp. Fuse			Norm. Fuse		
Weight Scheme		*uw*	*cfw*	*cw*	*uw*	*cfw*	*cw*
Prec.	5 docs	0.329	0.350	0.371	0.329	0.363	0.396
	10 docs	0.275	0.315	0.323	0.279	0.315	0.338
	15 docs	0.254	0.278	0.294	0.256	0.281	0.300
	20 docs	0.227	0.240	0.258	0.234	0.259	0.260
Av Precision		0.285	0.312	0.335	0.289	0.319	0.342

Table 69: VMR1b retrieval precision for data fusion combining 20K LVR and SI WS.

Weight Scheme		Simp. Fuse			Norm. Fuse		
		uw	*cfw*	*cw*	*uw*	*cfw*	*cw*
Prec.	5 docs	0.440	0.544	0.592	0.412	0.480	0.524
	10 docs	0.318	0.350	0.370	0.286	0.318	0.336
	15 docs	0.237	0.259	0.281	0.224	0.249	0.261
	20 docs	0.199	0.218	0.229	0.192	0.213	0.217
Av Precision		0.470	0.546	0.589	0.435	0.498	0.536

Table 70: VMR1a retrieval precision for data fusion combining 20K LVR and SI WS adapted using R13 data.

Weight Scheme		Simp. Fuse			Norm. Fuse		
		uw	*cfw*	*cw*	*uw*	*cfw*	*cw*
Prec.	5 docs	0.304	0.350	0.379	0.304	0.354	0.379
	10 docs	0.279	0.310	0.315	0.275	0.313	0.323
	15 docs	0.254	0.274	0.285	0.250	0.276	0.292
	20 docs	0.228	0.244	0.257	0.229	0.248	0.271
Av Precision		0.280	0.313	0.335	0.281	0.319	0.335

Table 71: VMR1b retrieval precision for data fusion combining 20K LVR and SI WS adapted using R13 data.

Weight Scheme		Simp. Fuse			Norm. Fuse		
		uw	*cfw*	*cw*	*uw*	*cfw*	*cw*
Prec.	5 docs	0.448	0.532	0.592	0.424	0.488	0.532
	10 docs	0.324	0.364	0.376	0.300	0.334	0.358
	15 docs	0.247	0.269	0.287	0.232	0.255	0.268
	20 docs	0.201	0.217	0.229	0.197	0.210	0.221
Av Precision		0.479	0.548	0.600	0.436	0.501	0.559

Table 72: VMR1a retrieval precision for data fusion combining 20K LVR and SI WS adapted using R75 data.

Weight Scheme		Simp. Fuse			Norm. Fuse		
		uw	*cfw*	*cw*	*uw*	*cfw*	*cw*
Prec.	5 docs	0.317	0.363	0.400	0.313	0.371	0.375
	10 docs	0.294	0.331	0.340	0.290	0.321	0.331
	15 docs	0.265	0.281	0.310	0.260	0.293	0.303
	20 docs	0.241	0.255	0.284	0.242	0.270	0.284
Av Precision		0.294	0.322	0.347	0.298	0.330	0.340

Table 73: VMR1b retrieval precision for data fusion combining 20K LVR and SI WS adapted using R75 data.

direct score combination LVR will dominate. With the much shorter VMR1b requests this effect is much reduced (perhaps absent) and WS is of considerably greater utility, as has already been observed.

Data Merging In data merging term hypotheses from LVR and WS are combined to produce a single document representation. WS hypotheses are thresholded based on their acoustic score, and all terms from both WS and LVR are suffix stripped before the individual evidence sources are merged. Tables 74 and 75 show retrieval performance results for data merging experiments using 20K WSJ LVR and SD WS. Two versions of this data merging procedure were carried out: adding only OOV WS words to the output of the LVR and combining all term hypotheses from both sources. For VMR1a retrieval performance varies very little with threshold, presumably because LVR dominates even with little thresholding for the reasons outlined in the previous section. For all data merging experiments results are shown for the WS acoustic threshold which gave overall optimal retrieval performance for the data merged system.

Tables 76 and 77 show the corresponding results for VMR1a and VMR1b using SI WS. Again, results are shown at the optimal threshold for merged LVR and WS document representations. Similarly, Tables 78 and 79 show data merging retrieval performance for SI WS adapted using R13 data, and Tables 80 and 81 show retrieval performance for SI WS adapted using R75 data.

Observations For both VMR1a and VMR1b it can be observed that data merging gives improved retrieval performance. Similarly it can be observed that combining all WS hypotheses with LVR output is for both collections better than merging only OOV terms with the LVR output. Once again it can be seen that retrieval performance involving WS hypotheses is strongly correlated to WS FOM.

Interestingly retrieval performance for both VMR1a and VMR1b is roughly comparable in percentage terms relative to text when using all WS terms in combination with LVR. In this arrangement the two indexing techniques appear to be complementary and best retrieval performance can be achieved regardless of request length.

5.4.2 Concluding Observations on LVR+WS Combined Methods

Figures 82 and 83 show a summary of retrieval performance for LVR and WS combinations for VMR1a and VMR1b respectively. Overall the following observations can be made (subject to the caveats about small collections). Data fusion is a very effective tool where LVR retrieval performance is insufficient to give near best overall retrieval performance. This is likely to be the case with VMR1b where the queries are relatively short.

Since in general the number of terms is not known in advance, the more reliable option is probably to implement data merging combining all WS terms. However, if short queries were observed and WS keywords had been well chosen for the domain, it might be better to employ normalised data fusion for these particular queries.

5.4.3 PLS + 20K LVR

This section describes experiments in combination of PLS and 20K WSJ LVR output. This system again has the open-vocabulary advantages of the PLS system described in

Weight Scheme		All Terms			OOV Terms		
		uw	*cfw*	*cw*	*uw*	*cfw*	*cw*
Prec.	5 docs	0.504	0.576	0.592	0.508	0.556	0.588
	10 docs	0.330	0.360	0.390	0.328	0.350	0.378
	15 docs	0.267	0.279	0.289	0.247	0.257	0.279
	20 docs	0.216	0.221	0.233	0.202	0.212	0.223
Av Precision		0.504	0.554	0.601	0.498	0.541	0.589

Table 74: VMR1a retrieval precision for data merging combining 20K LVR and SD WS.

Weight Scheme		All Terms			OOV Terms		
		uw	*cfw*	*cw*	*uw*	*cfw*	*cw*
Prec.	5 docs	0.317	0.304	0.342	0.296	0.300	0.329
	10 docs	0.260	0.273	0.285	0.240	0.260	0.273
	15 docs	0.236	0.242	0.268	0.222	0.239	0.268
	20 docs	0.220	0.216	0.243	0.203	0.217	0.241
Av Precision		0.268	0.279	0.316	0.250	0.270	0.294

Table 75: VMR1b retrieval precision for data merging combining 20K LVR and SD WS.

Weight Scheme		All Terms			OOV Terms		
		uw	*cfw*	*cw*	*uw*	*cfw*	*cw*
Prec.	5 docs	0.508	0.576	0.596	0.500	0.568	0.588
	10 docs	0.328	0.354	0.390	0.332	0.352	0.380
	15 docs	0.249	0.261	0.289	0.244	0.259	0.277
	20 docs	0.207	0.213	0.231	0.198	0.212	0.222
Av Precision		0.490	0.540	0.607	0.496	0.543	0.588

Table 76: VMR1a retrieval precision for data merging combining 20K LVR and SI WS.

Weight Scheme		All Terms			OOV Terms		
		uw	*cfw*	*cw*	*uw*	*cfw*	*cw*
Prec.	5 docs	0.296	0.279	0.346	0.292	0.308	0.329
	10 docs	0.235	0.254	0.267	0.235	0.267	0.265
	15 docs	0.221	0.236	0.260	0.226	0.239	0.260
	20 docs	0.198	0.201	0.241	0.204	0.206	0.233
Av Precision		0.248	0.265	0.306	0.250	0.272	0.290

Table 77: VMR1b retrieval precision for data merging combining 20K LVR and SI WS.

		All Terms			OOV Terms		
Weight Scheme		*uw*	*cfw*	*cw*	*uw*	*cfw*	*cw*
Prec.	5 docs	0.500	0.544	0.596	0.512	0.568	0.588
	10 docs	0.334	0.348	0.390	0.328	0.354	0.378
	15 docs	0.255	0.268	0.295	0.247	0.260	0.281
	20 docs	0.206	0.216	0.236	0.203	0.213	0.224
Av Precision		0.495	0.547	0.618	0.504	0.545	0.591

Table 78: VMR1a retrieval precision for data merging combining 20K LVR and SI WS adapted using R13 data.

		All Terms			OOV Terms		
Weight Scheme		*uw*	*cfw*	*cw*	*uw*	*cfw*	*cw*
Prec.	5 docs	0.317	0.317	0.338	0.296	0.317	0.329
	10 docs	0.263	0.271	0.277	0.240	0.275	0.279
	15 docs	0.236	0.238	0.269	0.229	0.253	0.265
	20 docs	0.214	0.217	0.250	0.206	0.221	0.241
Av Precision		0.268	0.281	0.316	0.253	0.278	0.297

Table 79: VMR1b retrieval precision for data merging combining 20K LVR and SI WS adapted using R13 data.

		All Terms			OOV Terms		
Weight Scheme		*uw*	*cfw*	*cw*	*uw*	*cfw*	*cw*
Prec.	5 docs	0.520	0.568	0.604	0.516	0.568	0.588
	10 docs	0.348	0.362	0.400	0.330	0.356	0.376
	15 docs	0.260	0.275	0.296	0.245	0.260	0.280
	20 docs	0.210	0.221	0.242	0.200	0.211	0.224
Av Precision		0.515	0.553	0.623	0.502	0.545	0.590

Table 80: VMR1a retrieval precision for data merging combining 20K LVR and SI WS adapted using R75 data.

		All Terms			OOV Terms		
Weight Scheme		*uw*	*cfw*	*cw*	*uw*	*cfw*	*cw*
Prec.	5 docs	0.317	0.304	0.350	0.296	0.308	0.338
	10 docs	0.252	0.273	0.288	0.238	0.269	0.267
	15 docs	0.232	0.244	0.275	0.233	0.246	0.265
	20 docs	0.217	0.213	0.251	0.208	0.210	0.239
Av Precision		0.265	0.278	0.324	0.255	0.276	0.295

Table 81: VMR1b retrieval precision for data merging combining 20K LVR and SI WS adapted using R75 data.

Weighting Scheme				Average Precision		
				uw	*cfw*	*cw*
Text		Avg. Prec.		0.600	0.671	0.718
		(relative)		100%	100%	100%
Spoken Documents	Data Fusion	LVR + SD WS	Simp. Fuse	79.2%	81.1%	81.9%
			Norm. Fuse	74.8%	75.3%	75.2%
		LVR + SI WS	Simp. Fuse	78.0%	80.2%	82.3%
			Norm. Fuse	71.0%	71.8%	72.6%
		LVR + SI WS R13	Simp. Fuse	78.3%	81.4%	82.0%
			Norm. Fuse	72.5%	74.2%	73.8%
		LVR + SI WS R75	Simp. Fuse	79.8%	81.7%	83.6%
			Norm. Fuse	72.7%	74.7%	77.9%
	Data Merging	LVR + SD WS	20K + all	84.0%	82.6%	83.7%
			20K + OOV	83.0%	80.6%	82.0%
		LVR + SI WS	20K + all	81.7%	80.5%	84.5%
			20K + OOV	82.7%	80.9%	81.9%
		LVR + SI WS R13	20K + all	82.5%	81.5%	86.1%
			20K + OOV	84.0%	81.2%	82.3%
		LVR + SI WS R75	20K + all	85.8%	82.4%	86.8%
			20K + OOV	83.7%	81.2%	82.2%

Table 82: Summary of VMR1a retrieval average precision for 20K LVR + WS.

Weighting Scheme				Average Precision		
				uw	*cfw*	*cw*
Text		Avg. Prec.		0.327	0.352	0.368
		(relative)		100%	100%	100%
Spoken Documents	Data Fusion	LVR + SD WS	Simp. Fuse	89.3%	90.3%	94.3%
			Norm. Fuse	90.2%	94.3%	95.7%
		LVR + SI WS	Simp. Fuse	87.2%	88.6%	91.0%
			Norm. Fuse	88.3%	90.6%	92.9%
		LVR + SI WS R13	Simp. Fuse	85.6%	88.9%	91.0%
			Norm. Fuse	85.9%	90.6%	91.0%
		LVR + SI WS R75	Simp. Fuse	89.9%	91.5%	94.3%
			Norm. Fuse	91.1%	93.8%	92.4%
	Data Merging	LVR + SD WS	20K + all	82.0%	79.3%	85.9%
			20K + OOV	76.5%	76.7%	79.9%
		LVR + SI WS	20K + all	75.8%	75.3%	83.2%
			20K + OOV	76.5%	77.3%	78.8%
		LVR + SI WS R13	20K + all	82.0%	79.8%	85.9%
			20K + OOV	77.4%	79.0%	80.7%
		LVR + SI WS R75	20K + all	81.0%	79.0%	88.0%
			20K + OOV	78.0%	78.4%	80.2%

Table 83: Summary of VMR1b retrieval average precision for 20K LVR + WS.

Section 5.2.2. Also since neither system PLS or LVR system is in any way specialised to the domain, this system represents completely domain independent indexing. Experimental results are shown for all 3 PLS acoustic model types.

Data Fusion Tables 84 and 85 show retrieval performance results from data fusion experiments combining 20K WSJ LVR with SD monophone PLS. Results are again shown for both simple fusion (where document matching scores are simply added together) and normalised fusion (where scores from each source are normalised relative to the highest score from that source before addition). After fusion the resulting new ranked list gives the results shown in the tables. For PLS, the *cw* list used in each case was obtained using message length measured in phones. (Note PLS data used for all Data Fusion experiments reported here were in each case those at the acoustic threshold which optimal retrieval performance for PLS on its own.) Tables 86 and 87 show retrieval performance for fusion using SI monophone PLS outputs and as before SI 20K WSJ LVR output. Tables 88 and 89 show fusion retrieval performance for SI biphone PLS and 20K WSJ LVR.

Observations Again as for data fusion of LVR and WS, VMR1a shows a preference for direct score fusion and VMR1b for the normalised form. Performance improvements for VMR1b are much more modest here than when using fusion with WS. In fact fusion retrieval performance for VMR1a is actually significantly degraded for most PLS models compared to LVR on its own. This is probably due to the poor performance of PLS in general relative to LVR for VMR1a. Only when using the SD PLS is there a modest improvement in retrieval performance for VMR1a. This result is slightly surprising since PLS actually performs better in isolation than WS for VMR1a, yet in combination for both collections is less effective than WS. A more useful comparison might be to look at the FOM results for WS and PLS shown in Tables 1 and 3 respectively. The FOM results for WS for the fixed keywords are considerably better for WS than for PLS. Although the rank of documents in a retrieval list produced using PLS indexing may itself be better than that for WS indexing, this tells only about the magnitude ordering of query-document matching scores in each list and nothing about the relative variation in the scores. The superior retrieval performance achieved using PLS is most likely to be due to its open search vocabulary since the reliability of individual term hypotheses will be lower for PLS. The matching scores in the PLS list are likely to contain much more "noise" derived from the presence of false alarms in the list. Hence when a PLS derived list is combined with an LVR derived list retrieval performance may actually be degraded relative to the LVR on its own. Note that the combined performance is always better than that using PLS on its own.

Data Merging Tables 90 and 91 show retrieval performance for data merging with SD monophone PLS and WSJ 20K LVR output. Very encouraging best retrieval performance here is well in excess of 90%. Tables 92 and 93 show retrieval performance for data merging with SI monophone PLS and WSJ 20K LVR output; and Tables 94 and 95 show retrieval performance for data merging with SI biphone PLS.

Observations Data merging of LVR and PLS shows a considerable improvement in performance over LVR or PLS alone for VMR1b. The difference here between including

Weight Scheme		Simp. Fuse			Norm. Fuse		
		uw	*cfw*	*cw*	*uw*	*cfw*	*cw*
Prec.	5 docs	0.452	0.520	0.572	0.440	0.512	0.564
	10 docs	0.318	0.342	0.390	0.316	0.348	0.392
	15 docs	0.245	0.260	0.288	0.248	0.260	0.291
	20 docs	0.199	0.212	0.233	0.202	0.213	0.236
Av Precision		0.473	0.538	0.585	0.465	0.538	0.584

Table 84: VMR1a retrieval precision for data fusion combining 20K LVR and SD monophone PLS.

Weight Scheme		Simp. Fuse			Norm. Fuse		
		uw	*cfw*	*cw*	*uw*	*cfw*	*cw*
Prec.	5 docs	0.333	0.325	0.321	0.329	0.304	0.354
	10 docs	0.288	0.275	0.296	0.277	0.277	0.313
	15 docs	0.236	0.257	0.260	0.234	0.253	0.265
	20 docs	0.209	0.225	0.227	0.222	0.224	0.235
Av Precision		0.266	0.298	0.316	0.271	0.297	0.323

Table 85: VMR1b retrieval precision for data fusion combining 20K LVR and SD monophone PLS.

Weight Scheme		Simp. Fuse			Norm. Fuse		
		uw	*cfw*	*cw*	*uw*	*cfw*	*cw*
Prec.	5 docs	0.388	0.488	0.544	0.376	0.456	0.536
	10 docs	0.284	0.318	0.342	0.282	0.302	0.344
	15 docs	0.212	0.251	0.257	0.216	0.241	0.255
	20 docs	0.170	0.198	0.211	0.171	0.195	0.205
Av Precision		0.409	0.489	0.552	0.402	0.468	0.539

Table 86: VMR1a retrieval precision for data fusion combining 20K LVR and SI monophone PLS.

Weight Scheme		Simp. Fuse			Norm. Fuse		
		uw	*cfw*	*cw*	*uw*	*cfw*	*cw*
Prec.	5 docs	0.304	0.317	0.304	0.283	0.286	0.300
	10 docs	0.235	0.256	0.267	0.231	0.240	0.252
	15 docs	0.208	0.225	0.242	0.204	0.208	0.238
	20 docs	0.190	0.212	0.232	0.192	0.202	0.222
Av Precision		0.235	0.261	0.284	0.232	0.244	0.268

Table 87: VMR1b retrieval precision for data fusion combining 20K LVR and SI monophone PLS.

		Simp. Fuse			Norm. Fuse		
Weight Scheme		*uw*	*cfw*	*cw*	*uw*	*cfw*	*cw*
Prec.	5 docs	0.420	0.528	0.548	0.428	0.532	0.552
	10 docs	0.294	0.340	0.376	0.298	0.340	0.376
	15 docs	0.237	0.253	0.279	0.240	0.251	0.275
	20 docs	0.193	0.209	0.222	0.197	0.211	0.227
Av Precision		0.444	0.518	0.566	0.444	0.520	0.562

Table 88: VMR1a retrieval precision for data fusion combining 20K LVR and SI biphone PLS.

		Simp. Fuse			Norm. Fuse		
Weight Scheme		*uw*	*cfw*	*cw*	*uw*	*cfw*	*cw*
Prec.	5 docs	0.317	0.325	0.317	0.321	0.308	0.308
	10 docs	0.263	0.287	0.281	0.273	0.292	0.290
	15 docs	0.233	0.251	0.260	0.232	0.247	0.263
	20 docs	0.212	0.220	0.226	0.213	0.216	0.225
Av Precision		0.255	0.286	0.301	0.258	0.281	0.300

Table 89: VMR1b retrieval precision for data fusion combining 20K LVR and SI biphone PLS.

		All Terms			OOV Terms		
Weight Scheme		*uw*	*cfw*	*cw*	*uw*	*cfw*	*cw*
Prec.	5 docs	0.496	0.560	0.616	0.488	0.560	0.608
	10 docs	0.330	0.360	0.406	0.334	0.356	0.396
	15 docs	0.244	0.271	0.295	0.252	0.271	0.291
	20 docs	0.203	0.216	0.240	0.203	0.219	0.232
Av Precision		0.491	0.561	0.634	0.493	0.560	0.626

Table 90: VMR1a retrieval precision for data merging combining 20K LVR and SD monophone PLS.

		All Terms			OOV Terms		
Weight Scheme		*uw*	*cfw*	*cw*	*uw*	*cfw*	*cw*
Prec.	5 docs	0.338	0.342	0.396	0.313	0.354	0.379
	10 docs	0.260	0.288	0.315	0.271	0.285	0.300
	15 docs	0.229	0.249	0.276	0.226	0.249	0.271
	20 docs	0.205	0.230	0.264	0.201	0.207	0.250
Av Precision		0.265	0.299	0.343	0.259	0.291	0.329

Table 91: VMR1b retrieval precision for data merging combining 20K LVR and SD monophone PLS.

Weight Scheme		All Terms			OOV Terms		
		uw	*cfw*	*cw*	*uw*	*cfw*	*cw*
Prec.	5 docs	0.476	0.548	0.568	0.452	0.536	0.584
	10 docs	0.312	0.344	0.384	0.310	0.350	0.378
	15 docs	0.227	0.252	0.281	0.236	0.265	0.283
	20 docs	0.188	0.205	0.222	0.192	0.216	0.230
Av Precision		0.448	0.512	0.584	0.463	0.531	0.589

Table 92: VMR1a retrieval precision for data merging combining 20K LVR and SI monophone PLS.

Weight Scheme		All Terms			OOV Terms		
		uw	*cfw*	*cw*	*uw*	*cfw*	*cw*
Prec.	5 docs	0.275	0.296	0.333	0.288	0.329	0.346
	10 docs	0.248	0.263	0.283	0.240	0.273	0.285
	15 docs	0.213	0.228	0.253	0.213	0.239	0.257
	20 docs	0.188	0.207	0.230	0.197	0.209	0.229
Av Precision		0.234	0.270	0.292	0.241	0.275	0.297

Table 93: VMR1b retrieval precision for data merging combining 20K LVR and SI monophone PLS.

Weight Scheme		All Terms			OOV Terms		
		uw	*cfw*	*cw*	*uw*	*cfw*	*cw*
Prec.	5 docs	0.476	0.552	0.580	0.464	0.544	0.604
	10 docs	0.318	0.338	0.380	0.328	0.350	0.378
	15 docs	0.236	0.251	0.289	0.252	0.261	0.283
	20 docs	0.192	0.210	0.232	0.199	0.214	0.229
Av Precision		0.481	0.530	0.592	0.472	0.530	0.601

Table 94: VMR1a retrieval precision for data merging combining 20K LVR and SI biphone PLS.

Weight Scheme		All Terms			OOV Terms		
		uw	*cfw*	*cw*	*uw*	*cfw*	*cw*
Prec.	5 docs	0.296	0.329	0.333	0.275	0.317	0.338
	10 docs	0.250	0.283	0.292	0.242	0.277	0.288
	15 docs	0.208	0.246	0.264	0.211	0.242	0.261
	20 docs	0.184	0.214	0.239	0.199	0.212	0.237
Av Precision		0.240	0.281	0.315	0.239	0.274	0.310

Table 95: VMR1b retrieval precision for data merging combining 20K LVR and SI biphone PLS.

all PLS or only OOV terms is much less marked than for LVR merging with WS. The choice of which scheme to use is apparently correlated to the quality of the PLS modelling. Thus inclusion of all PLS terms is seen to be a clear benefit for both VMR1a and VMR1b when using SD PLS whereas including only the OOV terms is actually slightly better in both cases when using SI monophone PLS. For SI biphone PLS there is no appreciable difference between the two merging schemes. In general this suggests that the better the quality of PLS modelling the more its evidence can usefully be exploited in data merging. Overall these results suggest that significant improvement in retrieval is possible if relatively modest further improvements in SI recognition could bring it closer to current SD recognition performance.

Figures 5 and 6 show average retrieval precision against acoustic score threshold for VMR1a and VMR1b respectively for 20K LVR, PLS with SI biphone models, and data merging using all term combination and merging only the OOV terms in each case using *cw* weighting. It can be seen from these figures that when using combination of all PLS terms with LVR the retrieval performance is, as expected, much more sensitive to the choice to threshold than when merging only the OOV terms.

5.4.4 Concluding Observations on LVR+PLS Combined Methods

Tables 96 and 97 show a summary of retrieval performance for VMR1a and VMR1b respectively for combination of 20K WSJ LVR and PLS. From these it can be seen that for LVR and PLS combination data merging is in generally preferable to data fusion, and that for data merging the overall quality of retrieval performance is strongly related to the quality of the acoustic PLS models. These results suggest that the investigation of unsupervised speaker-adaptation for SI biphone PLS could yield significant improvements in retrieval performance even if only modest gains in FOM were to be achieved. Also, interestingly it can be seen that in general *cw* weighting is less sensitive to variations in the quality of the acoustic modelling.

5.5 Results Summary

Tables 98 and 99 show overall summaries of experimental retrieval performance for the VMR project. These tables show only results for head-microphone modelling since no LVR and combination experiments were carried out using desk-microphone modelling. These tables give average precision for all systems relative to the ideal *open* text standard. As stated at the start of the experimental section, all these results must be treated with caution due to the small size of the test set. In particular, average precision, which is based on more information, may sharpen differences that are less apparent for the cutoff data, though the latter is likely to be more practically pertinent.

Bearing this in mind, we can observe that state-of-the-art LVR has utility for spoken document retrieval, in particular for queries with many terms. The 20K recogniser used here is now rather behind the true state-of-the-art in LVR, and hence if a current system with a 64K vocabulary and superior acoustic and language modelling were used improvements in retrieval performance would almost certainly appear.

Also, combination methods shown to be effective for text retrieval can also be used effectively in the speech domain. It should be noted however that the overall utility of this technique depends on the model quality of the individual components. Thus if one

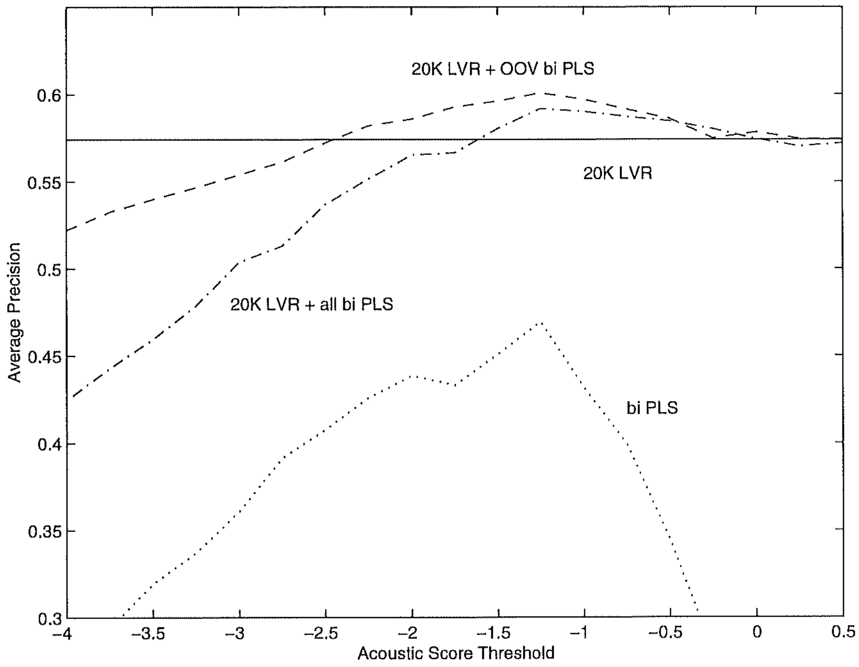

Figure 5: Retrieval average precision versus acoustic score threshold for 20K LVR and SI biphone data merging for VMR1a using *cw* weighting.

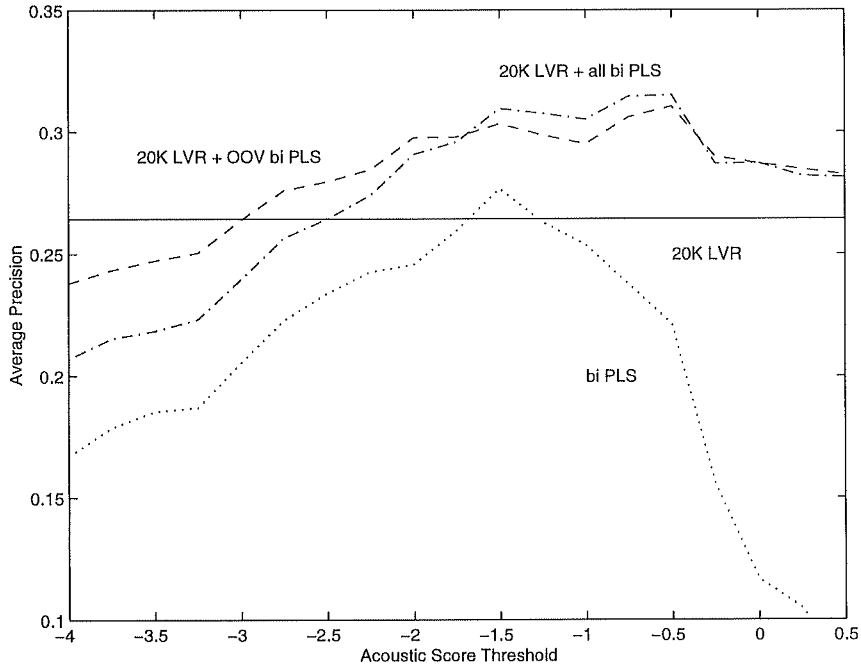

Figure 6: Retrieval average precision versus acoustic score threshold for 20K LVR and SI biphone data merging for VMR1b using *cw* weighting.

Weighting Scheme				Average Precision		
				uw	*cfw*	*cw*
Text		Avg. Prec.		0.600	0.671	0.718
		(relative)		100%	100%	100%
Spoken Documents	Data Fusion	LVR + SD PLS Monophone	Simp. Fuse	78.8%	80.2%	81.5%
			Norm. Fuse	77.5%	80.2%	81.3%
		LVR + SI PLS Monophone	Simp. Fuse	68.2%	72.9%	76.9%
			Norm. Fuse	67.0%	69.7%	75.1%
		LVR + SI PLS Biphone	Simp. Fuse	74.0%	77.2%	78.8%
			Norm. Fuse	74.0%	77.5%	78.3%
	Data Merging	LVR + SD PLS Monophone	20K + all	81.8%	83.6%	88.3%
			20K + OOV	82.2%	83.5%	87.2%
		LVR + SI PLS Monophone	20K + all	74.7%	76.3%	81.3%
			20K + OOV	77.2%	79.1%	82.0%
		LVR + SI PLS Biphone	20K + all	80.2%	79.0%	82.5%
			20K + OOV	78.7%	79.0%	83.7%

Table 96: Summary of VMR1a retrieval average precision for combination of 20K LVR and PLS.

Weighting Scheme				Average Precision		
				uw	*cfw*	*cw*
Text		Avg. Prec.		0.327	0.352	0.368
		(relative)		100%	100%	100%
Spoken Documents	Data Fusion	LVR + SD PLS Monophone	Simp. Fuse	81.3%	84.7%	85.9%
			Norm. Fuse	82.9%	84.4%	87.8%
		LVR + SI PLS Monophone	Simp. Fuse	71.9%	74.1%	77.2%
			Norm. Fuse	71.6%	69.3%	72.8%
		LVR + SI PLS Biphone	Simp. Fuse	78.0%	81.3%	81.8%
			Norm. Fuse	78.9%	79.8%	81.5%
	Data Merging	LVR + SD PLS Monophone	20K + all	81.0%	84.9%	93.2%
			20K + OOV	79.2%	82.7%	89.4%
		LVR + SI PLS Monophone	20K + all	71.6%	76.7%	79.3%
			20K + OOV	73.7%	78.1%	80.7%
		LVR + SI PLS Biphone	20K + all	73.4%	79.8%	85.6%
			20K + OOV	73.1%	77.8%	84.2%

Table 97: Summary of VMR1b retrieval average precision for combination of 20K LVR and PLS.

Weighting Scheme				Average Precision		
				uw	*cfw*	*cw*
Text	Full Vocab.	Avg. Prec.		0.600	0.671	0.718
		(relative)		100%	100%	100%
	20K Vocab.			96.0%	97.3%	97.9%
Spoken Documents	20K LVR			79.2%	77.9%	79.9%
	WS		SD	43.2%	44.0%	44.0%
			SI	40.2%	39.2%	41.8%
	PLS		SD Monophone	61.0%	63.6%	68.9%
			SI Monophone	42.2%	47.4%	54.3%
			SI Biphone	53.3%	61.3%	65.5%
	Data Fusion	LVR + SD PLS Monophone	Simp. Fuse	78.8%	80.2%	81.5%
			Norm. Fuse	77.5%	80.2%	81.3%
		LVR + SI PLS Monophone	Simp. Fuse	68.2%	72.9%	76.9%
			Norm. Fuse	67.0%	69.7%	75.1%
		LVR + SI PLS Biphone	Simp. Fuse	74.0%	77.2%	78.8%
			Norm. Fuse	74.0%	77.5%	78.3%
	Data Merging	LVR + SD PLS Monophone	20K + all	81.8%	83.6%	88.3%
			20K + OOV	82.2%	83.5%	87.2%
		LVR + SI PLS Monophone	20K + all	74.7%	76.3%	81.3%
			20K + OOV	77.2%	79.1%	82.0%
		LVR + SI PLS Biphone	20K + all	78.2%	77.3%	82.5%
			20K + OOV	78.7%	79.0%	83.7%

Table 98: Overall summary of VMR1a retrieval average precision values.

Weighting Scheme				Average Precision		
				uw	*cfw*	*cw*
Text		Avg. Prec.		0.327	0.352	0.368
		(relative)		100%	100%	100%
	20K Vocab.			91.4%	88.6%	88.3%
Spoken Documents	20K LVR			68.8%	69.9%	71.7%
	WS		SD	81.0%	88.6%	89.7%
			SI	76.8%	82.7%	81.8%
	PLS		SD Monophone	80.1%	81.0%	85.6%
			SI Monophone	53.2%	56.5%	60.3%
			SI Biphone	68.5%	74.4%	75.3%
	Data Fusion	LVR + SD PLS Monophone	Simp. Fuse	81.3%	84.7%	85.9%
			Norm. Fuse	82.9%	84.4%	87.8%
		LVR + SI PLS Monophone	Simp. Fuse	71.9%	74.1%	77.2%
			Norm. Fuse	71.6%	69.3%	72.8%
		LVR + SI PLS Biphone	Simp. Fuse	78.0%	81.3%	81.8%
			Norm. Fuse	78.9%	79.8%	81.5%
	Data Merging	LVR + SD PLS Monophone	20K + all	81.0%	84.9%	93.2%
			20K + OOV	79.2%	82.7%	89.4%
		LVR + SI PLS Monophone	20K + all	71.6%	76.7%	79.3%
			20K + OOV	73.7%	78.1%	80.7%
		LVR + SI PLS Biphone	20K + all	73.4%	79.8%	85.6%
			20K + OOV	73.1%	77.8%	84.2%

Table 99: Overall summary of VMR1b retrieval average precision.

Weighting Scheme				*uw*	*cfw*	*cw*	Table
Text	Open Vocab.			0.31	0.31	0.34	<< 25
	Keyword Vocab.			0.28	0.31	0.29	5
	20K Vocab.			0.28	0.28	0.29	45
Spoken Documents	20K LVR			0.21	0.24	0.25	<< 55
	WS		SD	0.27	0.32	0.30	7
			SI	0.24	0.28	0.31	<< 11
			SI + R75	0.26	0.30	0.32	17
	PLS		SD Monophone	0.27	0.25	0.29	35
			SI Monophone	0.16	0.18	0.20	<< 37
			SI Biphone	0.24	0.25	0.25	<< 39
	LVR + WS	Data Fusion	SD	0.29	0.33	0.34	67
			SI	0.28	0.32	0.34	<< 69
			SI + R75	0.29	0.32	0.33	73
		Data Merging	SD	0.26	0.27	0.29	75
			SI	0.24	0.25	0.27	<< 77
			SI + R75	0.25	0.27	0.29	81
	LVR + PLS	Data Fusion	SD Monophone	0.28	0.28	0.31	85
			SI Monophone	0.23	0.24	0.25	<< 87
			SI Biphone	0.27	0.29	0.29	<< 89
		Data Merging	SD Monophone	0.26	0.29	0.32	91
			SI Monophone	0.25	0.26	0.28	<< 93
			SI Biphone	0.25	0.28	0.29	<< 95

Table 100: Summary of VMR1b retrieval precision at document cutoff 10.

indexing system is significantly better than the other, combination can actually lead to a retrieval performance worse than the better system in isolation.

Perhaps the most important overall observation is that the best combination performance for SI biphone PLS and 20K LVR (a completely domain- and speaker-independent system) is clearly better than the performance of either method in isolation. Moreover, this combination produces retrieval performance of between 80% and 85% relative to text transcriptions for both VMR1a and VMR1b when using head-microphone data, though this is a rather more favourable environment that would occur in many practical situations.

5.5.1 Document Cutoff Results

Tables 98 and 99 give an overview of our test results using average precision. But it is also helpful, especially in relation to how users may view performance, to consider the picture of comparative performance given by document cutoff. Table 100 therefore summarises results for precision at document cutoff 10 for the more important of the two test collections, VMR1b. This cutoff is equivalent to the first page of output typically offered by operational services. As many of the later tests were restricted to head microphone ones, this table is only for head microphone. Where there are fine-grained choices for the different strategies, this table makes a consistent choice (in some cases this is the best option, in others there is in fact little difference). Thresholds, where relevant, were the best average precision ones, a reasonable approach for present purposes.

Thus for all cases where a stop list is applicable, this table shows the results for the van Rijsbergen list. For speaker adaptation with WS we use R75 data. For document length computation for PLS we use phones. For the fusion method of combination we use normalised scores, and for the merging method we merge for all terms. For clarity (and also because finer precision is hardly meaningful) we give values only to two decimal places. We mark the results of more general importance, i.e. those for speaker- independent speech recognition, with <<. We also give the source table citation at the right hand side.

Though, as repeated throughout this report, the collection was very small, there is sufficient consistency about some behaviours shown in Table 100 for it to be reasonable to take them as meaningful. Thus in general these figures show:

1. that for speech, term weighting is better than no weighting, with *cw* somewhat better than *cfw* .

2. that (in the speaker-independent case highlighted) for the three speech strategies PLS and LVR are not as good as WS. WS's merits are however attributable to the strong keyword tailoring of the collection, and the relatively inappropriate LVR modelling source. The LVR result in itself is also well below that for the absolute text reference case with open vocabulary. For the two combination methods the overall picture is rather varied; however it does appear that in general there is a gain from combination, with the fusion method working better than the merge one for LVR + WS but no obvious preference between the techniques for LVR + PLS. The potentially most useful combination from a practical point of view, namely LVR and PLS, does not however stand out here: it needs more study, for the various reasons mentioned.

3. that while speech performance is not as good as text, it is far from hopeless. Thus while the text reference case has precision 0.34 for *cw* ; while LVR has 0.25 and SD PLS has 0.29, in relation to the actual documents offered the user this is only a difference between 3.4 and 2.5 or 2.9 relevant documents out of 10 altogether.

6 Conclusions

The experiments reported here constitute only a first attack on open vocabulary retrieval for spoken documents. We have shown that it is possible to obtain speech retrieval performance, using open search terms, approaching that obtainable for text. Further, it appears that the combination of two recognition techniques can perform better than either alone, and indeed achieve an average retrieval precision for a SI system degraded by only 15% from the best achievable text retrieval. This difference can clearly be reduced by further improvements in speech recognition. Thus we have already found, for the SD case, that data combination retrieval performance using the current 20K LVR and PLS but with the SD monophones is only degraded by around 10% compared to text.

Fortunately for those concerned with spoken document retrieval, performance will continue to get better as the underlying speech recognition technology is improved. More sophisticated and efficient decoders mean that larger vocabularies may be used, which should reduce the OOV problem. Improvements in speech recognition can only benefit spoken document retrieval.

7 Suggestions for Further Work

The specific work reported in this report needs to be followed up in several ways. Primarily, it is essential to conduct retrieval tests on a larger scale, and with this in mind we have begun work on data capture and system development for television newscast retrieval [Brown et al., 1995].

At the same time, the approaches we have described must be developed to support a near real-time system. Even though expensive recognition is done offline, issues of storage and search efficiency must be addressed to yield a practical system. This is necessary both for larger-scale experiments and for operational use.

While improvements in speech recognition, as mentioned in the preceding section, will lead to a general improvement in spoken document retrieval, it is worth considering specific work which might be carried out for retrieval applications. For example, the results given in this report indicate that LVR and PLS in combination can lead to improved retrieval performance. However, to date these content discovery methods are used entirely independently in the indexing process. It is reasonable to suppose that further improvement might be possible if there was some interaction between these processes.

It is evident that when working with phones, recognition errors make it necessary to use a phone lattice rather than a single best phone transcription. When this is done it is also necessary to select an appropriate acoustic threshold on hypotheses as a compromise between insertions and deletions. With whole word recognition with LVR, however, we have so far used only a single best transcription. We have observed that for this LVR system, as we would anticipate, while it is prone to misses there are very few false alarms. It would seem an obvious extension of this existing work to consider the use of word lattices. As with phone lattices the deeper the lattice, the more false alarms we would observe; in this system parameters could expect be investigated to trade off between false alarms and misses. The key measure in this choice of parameters would be maximisation of the document retrieval performance.

An established technique in text information retrieval is *relevance feedback* where a search request may be iteratively modified based on relevance decisions by the user on documents already retrieved. Investigation of relevance feedback in spoken document retrieval is another obvious extension of the information retrieval research carried out within the VMR project.

8 Acknowledgements

We are grateful to Steve Walker and Steve Robertson of City University for supplying the stop word lists used in this work.

References

[Belkin et al., 1995] Belkin, N. J., Kantor, P., Fox, E. A., and Shaw, J. A. (1995). Combining the evidence of multiple query representations for information retrieval. *Information Processing and Management*, 31(3):431–448.

[Brown et al., 1995] Brown, M. G., Foote, J. T., Jones, G. J. F., Jones, K. S., and Young, S. J. (1995). Automatic content-based retrieval of broadcast news. In *Proc. ACM Multimedia 95*, pages 35–43, San Francisco. ACM.

[Foote et al., 1994] Foote, J. T., Jones, G. J. F., Spärck Jones, K., and Young, S. J. (1994). Video Mail Retrieval using Voice: Report on whole word based keyword spotting. Technical report, Cambridge University Engineering Department.

[Foote et al., 1995] Foote, J. T., Jones, G. J. F., Spärck Jones, K., and Young, S. J. (1995). Talker-independent keyword spotting for information retrieval. In *Proc. Eurospeech 95*, volume 3, pages 2145–2148, Madrid. ESCA.

[Foote et al., 1996] Foote, J. T., Jones, G. J. F., Spärck Jones, K., and Young, S. J. (1996). Video Mail Retrieval using Voice: Report on phone lattice spotting. Technical report, Cambridge University Engineering Department.

[Foote et al., 1997] Foote, J. T., Young, S. J., Jones, G. J. F., and Spärck Jones, K. (1997). Unconstrained Keyword Spotting using Phone Lattices. *Computer Speech and Language*. In press.

[Hopper et al., 1993] Hopper, A., Spärck Jones, K., and Young, S. J. (1993). VMR Video Mail Retrieval using Voice. Research Proposal: Olivetti Research Limited, Cambridge University Computer Laboratory & Cambridge University Engineering Department.

[James and Young, 1994] James, D. A. and Young, S. J. (1994). A fast lattice-based approach to vocabulary independent wordspotting. In *Proc. ICASSP 94*, volume I, pages 377–380, Adelaide. IEEE.

[Jeanrenaud et al., 1995] Jeanrenaud, P., Eide, E., Chaudhari, U., McDonough, J., Ng, K., Siu, M., and Gish, H. (1995). Reducing word error rate on conversational speech from the Switchboard corpus. In *Proc. ICASSP 95*, pages 53–56, Detroit. IEEE.

[Jones et al., 1994] Jones, G. J. F., Foote, J. T., Spärck Jones, K., and Young, S. J. (1994). VMR report on keyword definition and data collection. Technical Report 335, Cambridge University Computer Laboratory.

[Jones et al., 1995b] Jones, G. J. F., Foote, J. T., Spärck Jones, K., and Young, S. J. (1995b). Video Mail Retrieval: the effect of word spotting accuracy on precision. In *Proc. ICASSP 95*, volume I, pages 309–312, Detroit. IEEE.

[Jones et al., 1995a] Jones, G. J. F., Foote, J. T., Spärck Jones, K., and Young, S. J. (1995a). Video Mail Retrieval using voice: An overview of the stage 2 system. In van Rijsbergen, C. J., editor, *Proceedings of the MIRO workshop*, University of Glasgow. Springer-Verlag.

[Jones et al., 1996a] Jones, G. J. F., Foote, J. T., Spärck Jones, K., and Young, S. J. (1996a). Video Mail Retrieval using voice: Report on collection of naturalistic requests and relevance assessments. Technical Report 402, Cambridge University Computer Laboratory.

[Jones et al., 1996b] Jones, G. J. F., Foote, J. T., Spärck Jones, K., and Young, S. J. (1996b). VMR report on large vocabulary speech recognition. Technical report, Cambridge University Engineering Department.

[Leggetter and Woodland, 1995] Leggetter, C. J. and Woodland, P. C. (1995). Flexible speaker adaptation for large vocabulary speech recognition. In *Proc. Eurospeech 95*. ESCA.

[McDonough et al., 1994] McDonough, J., Ng, K., Jeanrenaud, P., Gish, H., and Rohlicek, J. R. (1994). Approaches to topic identification on the switchboard corpus. In *Proc. ICASSP 94*, volume I, pages 385–388, Adelaide. IEEE.

[Peskin et al., 1996] Peskin, B., Connolly, S., Gillick, L., Lowe, S., MaAllaster, D., Nagesha, V., van Mulbreg, P., and Wegmann, S. (1996). Improvements in switchboard recognition and topic identification. In *Proc. ICASSP 96*, volume I, pages 303–306, Atlanta. IEEE.

[Porter, 1980] Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

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

[Robertson and Spärck Jones, 1994] Robertson, S. E. and Spärck Jones, K. (1994). Simple, proven approaches to text retrieval. Technical report, Cambridge University Computer Laboratory.

[Robertson and Walker, 1994] Robertson, S. E. and Walker, S. (1994). Some simple effective approximations to the 2–Poisson model for probabililtic weighted retrieval. In *Proc. SIGIR 94*, pages 232–241, Dublin. ACM.

[Robertson et al., 1995] Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., and Gatford, M. (1995). Okapi at TREC-3. In Harman, D. K., editor, *The Third Text REtrieval Conference (TREC-3)*, pages 109–126.

[Robinson et al., 1995] Robinson, T., Fransen, J., Pye, D., Foote, J., and Renals, S. (1995). WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition. In *Proc. ICASSP 95*, pages 81–84, Detroit. IEEE.

[Rose, 1991] Rose, R. C. (1991). Techniques for information retrieval from speech messages. *Lincoln Laboratory Journal*, 4(1):45–60.

[Spärck Jones et al., 1995] Spärck Jones, K., Foote, J. T., Jones, G. J. F., and Young, S. J. (1995). Spoken document retrieval — a multimedia tool. In *Fourth Annual Symposium on Document Analysis and Information Retrieval*, pages 1–11, Las Vegas.

[Spärck Jones et al., 1996] Spärck Jones, K., Jones, G. J. F., Foote, J. T., and Young, S. J. (1996). Experiments in spoken document retrieval. *Information Processing and Management*, 32(4):399–417.

[van Rijsbergen, 1979] van Rijsbergen, C. J. (1979). *Information Retrieval.* Butterworths, London, 2nd edition.

[Wechsler and Schäuble, 1995] Wechsler, M. and Schäuble, P. (1995). Speech retrieval based on automatic indexing. In van Rijsbergen, C. J., editor, *Proceedings of the MIRO Workshop*, University of Glasgow.

[Wright et al., 1995] Wright, J. H., Carey, M. J., and Parris, E. S. (1995). Improved topic spotting through statistical modelling of keyword dependencies. In *Proc. ICASSP 95*, pages 313–316, Detroit, MI. IEEE.

[Young et al., 1994] Young, S. J., Odell, J. J., and Woodland, P. C. (1994). Tree-based state tying for high accuracy acoustic modelling. In *Proc. ARPA Spoken Language Technology Workshop*, Plainsboro, NJ.

[Young et al., 1993] Young, S. J., Woodland, P. C., and Byrne, W. J. (1993). *HTK: Hidden Markov Model Toolkit V1.5.* Entropic Research Laboratories, Inc., 600 Pennsylvania Ave. SE, Suite 202, Washington, DC 20003 USA.