Technical Report

UCAM-CL-TR-366
ISSN 1476-2986

Number 366

UNIVERSITY OF CAMBRIDGE

Computer Laboratory

Retrieving spoken documents: VMR Project experiments

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

May 1995

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
DOI *https://doi.org/10.48456/tr-366*

Retrieving spoken documents: VMR Project experiments *

K. Sparck Jones[†], G.J.F. Jones[†‡], J.T. Foote[‡] & 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

May 1995

Abstract

This paper describes initial work on an application for the retrieval of spoken documents in multimedia systems. Speech documents pose a particular problem for retrieval since the contents are unknown. The VMR project seeks to address this problem for a video mail application by combining state of the art speech recognition with established document retrieval technologies to provide an effective and efficient retrieval tool. Experiments with a small spoken message collection show that retrieval precision for the spoken file can reach 90% of that obtained when the same file is used, as a benchmark, in text transcription form.

*This project is supported by the UK DTI Grant IED4/1/5804 and SERC (now EPSRC) Grant GR/H87629. Olivetti Research Limited is an industrial partner of the VMR project. The VMR corpus used for this work will be available for public distribution in the near future.

1 Introduction

This paper presents initial work on a novel multi-media retrieval application. The Video Mail Retrieval (VMR) project seeks to combine state of the art speech recognition and document retrieval technologies for spoken message retrieval, envisaged as one function among many provided on a workstation equipped with multimedia video facilities.

The paper discusses the problems involved, the strategies being deployed to overcome these, the initial system implementation, and the design and results of our first retrieval tests. Our claims are on the one hand, that the straightforward probabilistic methods that have been established for text retrieval can be naturally extended to the speech case; and on the other, that current speech recognition technology can support adequate message retrieval performance, even if this is not as accurate as that obtained with written text. Our further claim is that spoken message retrieval can be conveniently implemented to enhance the capabilities of an office computer system with video facilities. The work and results reported here provide data to support these claims, though as this is research in progress we have much more development and testing to do. Specifically, our experiments constitute the first serious tests of spoken document retrieval combining modern recognition and retrieval technologies, even if they have so far been on a relatively small scale.

Section 2 presents the VMR Project background to provide a context for the subsequent discussion of the distinctive problems to be overcome in speech retrieval and to motivate our own approach. Section 3 summarises the specific technical problems to be tackled in spoken document retrieval. The overall strategy we have adopted for the project is outlined in Section 4, which is followed by Sections 5 to 7 describing our work so far in detail, covering data provision in 5, speech processing in 6 and retrieval testing in 7, concluding with an assessment of what we have done to date. The final section, 8, comments on current work and summarises our planned future research. The paper includes some previously published material, reproduced here for comparative purposes or completeness.

2 Background

2.1 The Pandora System

The VMR Project stems from the Pandora multimedia system developed jointly by Olivetti Research Limited (ORL), Cambridge and the Computer Laboratory. Pandora provided facilities for both video interaction between users and video mail facilities. In practical operation it became evident that large volumes of material were accumulating for which some content-based search apparatus was required, ie one exploiting recorded sessions and message bodies, and not confined to normal administrative *header* data (eg participants' names, dates etc) or the very limited content indicators of header 'subject' lines. The same situation has already been observed with email, the need for a proper search apparatus becoming more important with the passage of time, and with the use of the Pandora video system for substantive matters and as a major channel for report, discussion and decision.

The VMR Project described here was intended to attack the most obviously tractable, and also potentially probably most useful, Pandora retrieval task, namely mail retrieval on the spoken record, using the image information only in supporting display. From the

speech recognition point of view the messages are clearly more manageable than dialogues, while from a content retrieval point of view they are often more concentrated and definite. At the same time, while images may provide rich information, direct image retrieval by content specification is not well developed, and the images in Pandora are predominantly only those of individual speakers belonging to a limited community that are not necessarily sufficient to flag or distinguish message content further than their names in headers already do.

The Pandora system is now being replaced at ORL by the Medusa system; this is intended to embody many specific improvements over Pandora, notably in the quality of its audio facilities; but does not affect the fundamental VMR project work, other than in relation to data for testing and evaluation, as described later.

2.2 Related retrieval applications

The project naturally applies retrieval techniques suited to full text; however, some specific text retrieval applications, most obviously to email, are more germane than others.

2.2.1 Email

Video messages, especially where the speaker naturally adopts a dictation-like style, may resemble conventional abstracts in length, content density and presentational coherence. However, they may be much more informal and incomplete, especially where they are elements of an extended exchange of messages. Operational systems are already available for searching large email files (eg Usenet (Burrows 1991)), and search facilities are proliferating in an adhoc way for other data sources including email, notably for Internet, though the retrieval techniques used may be rudimentary and far from best established practice, and performance has not been properly assessed. The email case can, however, provide experience in effective ways of combining search keys referring to message headers with ones referring to message bodies that is more relevant to the VMR case, because of the particular descriptive characteristics of email headers, than ways of combining bibliographic header and text keys in conventional document retrieval. One important difference between email and video mail, on the other hand, is that while the replication of previous message material in new messages is becoming common with email, this does not now apply to video mail and might not become common even if technologically feasible, because of increasing the time taken to 'read' messages.

2.3 Retrieval involving speech

Spoken access to text files is a developing area (Kupiec, Kimber & Balasubramanian 1994), but in general transcribing spoken queries for text file searching is much less challenging than speech file searching. For instance, it is more reasonable to expect the searching user to speak clearly and accept verification of search terms, and to participate in query development. Spoken query work is pertinent to VMR, but the much more important challenge in the VMR case is clearly that of dealing with the spoken document files.

2.3.1 Access to speech files

The current dominant applications for speech recognition, illustrated by the (D)ARPA initiatives on Continuous Speech Recognition (CSR) and Spoken Language Systems (SLS), have helped to raise speech processing standards but are not directly relevant, in terms of the system purposes involved, to VMR. One of the tasks involved, the transcription of continuous read speech (known as the Wall Street Journal recognition task), nevertheless bears on VMR both both in supplying generic performance data for word identification in continuous speech and in acting as a testbed for the specific speech recognition system (Young, Woodland & Byrne 1993) being used for VMR. The kind of performance measure for recognition exemplified by the ARPA Word Error Rate (Woodland et al 1995) is not, however, especially suited to the retrieval context, since it is essentially about the successful recognition of *all* the words in running speech, treating them as equally required and important; it does not concentrate on recognition behaviour for particular content-bearing words, or even all such words as opposed to function words which are not useful from a retrieval perspective. Furthermore, much of this work is based on read speech. Spontaneous speech recognition is much harder and Word Error Rates are typically much higher (Young, Woodland & Byrne 1994).

2.3.2 Word spotting

An alternative recognition requirement is for *word spotting*, focusing only on the recognition of important content words. Video mail messages pose a much more difficult recognition problem than material that has been read, since the speech is often informally structured (from a linguistic point of view) and there is much greater variation in utterance style over time. For these reasons speech recognition in the VMR project is most usefully treated as a word spotting enterprise. This does however involve more than a method of identifying just required words: since a speech recogniser must hypothesise the occurrence of some speech event at all times, word spotters usually incorporate additional models to cover periods of silence along with a "filler" model for recognising all non-keyword acoustic events.

Word spotting has long been an intelligence agency concern, but it is not made easy to know what has been achieved here, and how specifically pertinent it is to VMR. Thus, the files may be different, involving both low, telephone, quality acoustic data, and conversations as documents; equally, both requests and relevance criteria may be quite different from those for conventional document retrieval by topics, eg seek only explicit occurrences of specific names; and the task may be routing or filtering rather than one-off searching.

Some work on word spotting has been reported, for example by Rose (1991), Wilcox & Bush (1992) and McDonough et al. (1994), though for varied application tasks, in some cases with similarities to retrieval, eg message categorisation (Rose 1991, McDonough 1994), but in others for quite different purposes (Wilcox & Bush 1992). However even with good speech performance, the categorisation work described used classes that were much coarser than those normally defined by message retrieval requests, so this work has little direct bearing on our own.

The forms of performance measure used in this work on word spotting are however far more appropriate than the Word Error Rate. They exploit notions, with respect to matching for sought words, of *hit*, *miss* and *false alarm* which have obvious analogues in

document matching, and which reflect the emphasis in word spotting on the identification of particular words within the speech stream. However, a measure like the False Alarm (FA) Rate, used in conjunction with the percentage of hits to provide a Figure of Merit (FOM) characterising performance, is still somewhat remote and needs to be related to the realities of search term occurrence in documents.

2.3.3 Conventional retrieval

Little work has been done in this area, and what has been done has been limited in scope or evaluation. Schäuble and others (Glavitsch & Schäuble 1992, Schäuble & Glavitsch 1994, Glavitsch, Schäuble & Wechsler 1994) have proposed a system for speech retrieval based on predefined acoustic units, and have considered the effect of term occurrence errors of semantic and acoustic origin, but only by simulation. In particular, while Glavitsch and Schäuble and their colleagues show that retrieval performance can hold up reasonably well under some noise (as is required for actual spoken document retrieval), their 'consonant/vowel' index key model drawn from text does not carry over in an unproblematic way to speech: the indexing keys are so brief they are likely to be very unreliable, as some studies conducted by James in Cambridge (James 1995) suggest; and their simulation of recognition errors is not acoustically motivated, no account is taken of the important matter of substitution errors, and insertion and deletion errors are assumed to be simply distributed over all documents. Finally, their tests with some classical IR test collections have used written documents which do not necessarily have the same linguistic and discourse properties as spoken ones.

Genuine spoken document retrieval experiments have, however, been recently carried out by James (James 1995), using short news stories and semi-realistic prompted requests. James applied conventional term weighting in a straightforward way in conjunction with indexing by phones, and explored a range of alternative strategies for term recognition. He obtained best performance for a combination of word recognition for a standing vocabulary, using triphones and a bigram language model, with matching via a biphone lattice for out-of-vocabulary words. This gave average precision performance, for forty requests and over three hundred documents, reaching 90% of that obtainable with the same methods for text versions of the data.

James's experiments are valuable in suggesting that speech processing and information retrieval technologies can be effectively combined. However the data used, professionally spoken and consisting of short items with a high information content, cannot be taken as typical of many applications. The tests described below were motivated by the need to develop methods suited to more demanding conditions, especially in relation to the discourse properties of file documents; and they also exploited consistent methods of term weighting.

3 Problems in spoken document retrieval

The VMR Project is applying well-understood probabilistic retrieval techniques (Robertson & Sparck Jones 1994) and equally well-established speech recognition techniques (Rabiner 1989), Thus the retrieval methods use simple natural language units (stems, words, phrases) in coordination, with statistically-based weighting, to produce a ranked search

output. The speech techniques rely on hidden Markov models (HMMs), where the identification of any particular speech unit (phone, word, word string) is computed using a sequence of transition probabilities for items given their predecessors, based on data from a prior training sample. For any given unit, identification takes the form of an acoustic score, or *a-score*. However, in combining these retrieval and speech techniques for spoken document retrieval there are a number of key problems.

These problems are primarily those arising from the use of speech as such, and secondarily those associated with the particular kind of documents involved, but both have implications for any attempt to apply the retrieval methods hitherto successfully used for written text. Thus the project has to focus on the interaction effects of speech and retrieval factors, seeking retrieval offsets to speech challenges. But this is made difficult by the current lack of experience of speech retrieval and by the difficulties of getting performance reference data for large speech files. The problems we have are as follows.

3.1 Term constraints

For individual terms, the project has to tackle all the usual problems of word recognition in continuous speech: compared with the text case there is always, in the most fundamental way, some uncertainty, reflected in a-score values, about whether a putative word is actually present in the speech at that time. Lexical category and sense ambiguity are familiar in text retrieval. But while such ambiguities may be resolved in the speech case by different pronunciation, new ambiguities appear through homophones and, especially, alternative word boundary allocations (eg 'Hello Kate' vs 'locate'). There is a further problem with stemming, since in the spoken case the same stem may be differently pronounced depending on its full form context. A more fundamental problem is that word spotting on shorter words is inherently more unreliable and hence generalising search terms by suffix stripping may actually degrade word spotting reliability.

The natural strategies for dealing with the uncertainty and ambiguity problems are the query formulation ones already applied to reduce ambiguity in the text case, namely increasing the number of search terms, exploiting redundancy to secure multi-term matches. Thus these techniques can be readily combined with the use of thresholds on acoustic scores, which are useful as a means of avoiding false alarms in word matching. However because thresholds also mean that true hits that happen to have low a-scores are lost, it is necessary to study the effects of threshold setting on trade-offs between hits, misses and false alarms and, more importantly, to see how the potentially valuable data provided by actual a-scores could be exploited. This is a complicated matter both because the occurrences of a word naturally have different a-scores and because ways have to be found of combining numerical term values given by the distinct procedures used for acoustic processing and term weighting. Thus, while the speech recognition and document retrieval approaches we are using are both probabilistic, it is still necessary to find proper ways of combining speech recognition a-scores with the usual types of term weights in order to derive single, meaningful index term values and consequent query-document matching scores.

3.1.1 Term data limitations

Spoken document files consist of digitised acoustic records, ie contain sound streams, not explicit word sequences; and because speech recognition will be far from perfect in the kinds of context with which we are concerned, while document retrieval is essentially a bulk task, we cannot rely on being able to gather any very reliable, detailed or exhaustive data about the occurrences of words in the document files from direct operations on these files. This affects design and testing and, critically, actual retrieval options.

Even for test purposes it is impracticable to gather any significant amount of guiding or reference information about word behaviour via transcription to text or by listening to file material. But more importantly, there is a fundamental problem for operational speech retrieval in that it cannot, in general, be assumed that there is a list of all the words used in the file, with accompanying incidence data. Even if there is an adequate reference lexicon supplying word pronunciations, and it is feasible in principle to exploit this to find what words may occur in the speech files, it is not possible in practice to extract all the words from the files and build comprehensive indexes independent of the selectivity provided by working only with specific search terms. Retrieval strategies have therefore to allow for eg filtering and multiple processing cycles to gather and apply occurrence data for weighting, even when working just with search terms from input requests.

3.2 Message constraints

The specific ways in which term uncertainty and ambiguity are managed naturally depend on the properties of queries and documents eg their length, content density etc, as this affects the chances of matching on single occurrences of search terms. The VMR messages also present particular challenges to normal document retrieval protocols through their informality and/or brevity, and while there is currently, as noted, growing interest in email retrieval there is no solid data about file behaviour and search performance for email under established retrieval methods. Thus there may be difficulties with email at the individual document level eg because messages are shorter than full papers but less concentrated than abstracts; and at the file level because there are many duplicate documents due to forwarding and encapsulation. Message and file problems come together in message chains embedding minimal individual messages like "You're right: OK - Jack". With speech mail individual documents may be more informal, dilute or cryptic than written ones, and while less likely to involve extensive repetitive quoting, perhaps be repetitive in other ways.

There are clearly serious implications for system design, testing and operation in the lack of information about words and hence about important properties of documents. It is indeed possible to obtain information about the gross properties of a file, eg about average message length and variation in length, and also, from sampling and listening, about the flavour and style of the documents. But implementing a retrieval system based on statistical techniques, and choosing appropriate strategies for query formation, is a difficult matter when information about individual words is not readily available, or is perhaps in practice not available at all. The lack of direct information about words also affects dependent information both about internal message characteristics like content word density and about external message relationships, and also information about the content properties of the message set as a whole.

There are particular difficulties with some specific techniques found helpful in the text

case, notably query expansion in relevance feedback. It may not be possible to identify other retrieved message words to offer the user as candidate additions to the query or to add to it automatically. The user will have to process the retrieved output more minutely than in the written case, and explicit file processing will be needed to gather frequency data to assign relevance weights to new terms for further searching, again implying much more complex operations than with text systems.

3.3 User interface constraints

Speech retrieval also presents extra challenges for interface design to promote effective user searching. The most obvious point is that listening to speech takes far more time than scanning text, so where eg twenty abstracts might take 30 seconds reading, twenty messages of similar word length could take 15 minutes listening. It may be possible to supply (playback) short extracts eg parts of messages where search terms are concentrated, but these may not be very informative if there are few or common search terms, or the extract is short. It is independently desirable to display other document information, notably message headers, and in principle in the VMR context displaying the accompanying sender image may be helpful. The retrieval interface has to be developed to allow the user to exploit the various types of information associated with messages in a convenient way, both by combining these for searching, in *hybrid* searches, or by pursuing distinct types of header, as in chain following. As with email, the retrieval interface in the strict sense has also to be envisaged as just one element of the user's entire workstation apparatus. We are currently experimenting with a visual display of file contents as a horizontal bar showing putative positions of search terms which enables the user to manually select potentially interesting sections to play back.

4 Project strategy

Our project strategy has been designed to respond to these problems, but has also been influenced by local factors. Thus the project has three phases, intended both to explore speech retrieval under progressively less restricted and more realistic conditions, and to embed speech retrieval capabilities in a habitable user interface.

In the first stage, searching depends on a fixed keyword vocabulary, and assumes a closed speaker community responsible for the file messages. The retrieval strategy is a simple one and interface facilities are also limited. This first phase specification allows word recognition under the most favourable conditions, ie where the speech recogniser can be trained to focus on a known word set as used by a known speaker set in a noiseless recording environment. This is clearly unrealistic in the long run, but provides a benchmark for performance in later, less favourable conditions.

The second stage design allows for an open speaker community, and the third for an open search term vocabulary as well. These later stages also cover more sophisticated retrieval techniques for terms and more flexible facilities for hybrid retrieval, and the development of the general user interface. The successive stages were also planned to involve tests with progressively more demanding document collections, and with more varied user needs and queries.

We have, however, had to deal with an unexpected problem, the lack of real Pandora message data, as a source both of general guidance and of actual test document collections. Pandora messages were recorded with relatively low quality microphones, making the audio recordings too poor for any form of automated speech recognition. The existing Pandora file content also contains confidential material and is rather old, making it difficult to develop search queries and assess output. The successor Medusa system has much improved (and usable) audio quality, but has so far not been deployed in a routine way involving a sizeable user community and naturally generating a message archive. Thus while Medusa can be used for interface development and user trials for retrieval system convenience etc., we have had to engage in a serious collection construction enterprise for our initial retrieval test data, and we expect to use other non-Medusa material for future experiments, as described later.

5 Data provision

Our first message set, VMR1[1], was designed to meet various criteria arising from the requirements of both the document retrieval component of the system and the speech recognition element. From the document retrieval perspective the database had to consist of messages with the same general properties as could be expected in Pandora/Medusa-type installations. But in order to meet the specification for Stage 1 of the project, it also had to consist of messages making natural use of a set of fixed search keywords, while at the same time forming a corpus with the kinds of message similarities and differences, that pose challenges for recall and precision, as could be deemed typical of expected VMR situations. Messages should also have similar acoustic properties to those found in an operational system and be of comparable length. At the same time, we could not rely on having many people to generate messages specifically for us, independent of any operational video system with a natural user community. We also wished to transcribe the message set to provide a reference base for testing, but could not carry out large-scale transcription since this is a comparatively expensive process. VMR1 is therefore a very small document collection from the retrieval point of view, serving primarily as a tool for development of our speech processing technology and general approach, but not allowing serious retrieval experiments.

Apart from messages, VMR1 had also to contain spoken material for use as training data for the speaker-dependent acoustic models in the word spotter. The appropriate training of acoustic model parameters is an important feature in the building of speech recognition systems and it is vital to have suitable and sufficient data to construct a high performance system.

VMR1 is fully described in Jones et al. (1994); the main points about it are as follows.

5.1 Message definition

While the structure of the message set is important if a small database is to be viable for retrieval research, it is also necessary to have messages that are natural in both content and speaking style. To meet the first requirement we sought messages on *topics* within a set of topic *categories*: each category has an associated set of *keywords* drawn from a

[1]also previously called Database 1 and sometimes referred to as a corpus

fixed keyword *vocabulary* from which search *terms* in Stage 1 must be taken. To meet the second requirement, since the messages were collected specially and not drawn from a natural mail community, we utilised prompting *scenarios*. These stimulated the speaker to talk on a topic within a category without constraining them to produce messages strictly tied to pre-specified topics. The scenarios also encouraged the speakers to use the keywords for the category, though keywords were not exclusive to categories and could well occur in messages in other categories. However as the scenarios were only prompts, messages for the same scenario could be on quite different individual topics within the same general content area. Overall, our approach led to messages that varied in their individual topics but that clustered round the prompting categories, while maintaining links, through broader relations between categories marked by shared keywords, with messages in other categories.

In addition, since the keyword vocabulary is not very large, we provided a set of *otherwords* as an further prompting and potential search vocabulary.

It should be noted that while our general notion of topic also applies to requests, the message database was not constructed with any test requests already in mind. The provision of test requests and relevance assessments is described in Section 7.

5.2 VMR1 database formation

We used a total keyword vocabulary of 35 words, along with 31 related otherwords, and 10 broad subject categories. We collected 300 messages, taken as the maximum practicable for transcription and as the minimum sufficient for initial testing of our speech retrieval processes in conditions that share critical properties with real ones, though we of course do not regard a set of 300 documents as a proper IR test collection. Thus all the choices of keywords, topics and categories were motivated by the need for a 'representatively' naturalistic message set, and the same considerations, applying both to the characteristics of their messages and to their speech, determined the number of speakers used and number of messages per speaker.

Thus we sought 20 messages each from a subset of 4 categories from the 10 available, with 15 speakers overall, so that for any one category there were messages from 6 speakers. The assignment of speakers to categories was randomised, and the actual data collection protocol was designed to encourage an even distribution of messages across the scenarios within a category for each speaker, although this could not be enforced.

The *prompt* for each spontaneous message consisted of the scenario and the keywords and otherwords for the category. Speakers were asked to favour the use of the listed keywords and otherwords, but not at the expense of construction of realistic messages. They were also not restricted to the keywords precisely as shown to them but could use them in variant *word forms*: for example, the keyword *mail* might be used in the forms *mailed*, *mails* or *mailing*. (The word spotter should pick up the common stem as long as there is not too much pronounciational variation, and so get correct hits on the keyword). The speakers were not shown a complete list of the keywords available, but only those relevant to the current category.

Speakers were assigned to categories about which they were deemed knowledgeable, to encourage plausible messages; but to avoid sequencing effects for categories and scenarios both category and scenario orders were distributed per speaker on a Latin square basis

(Tague 1981).

Training data, in the form of material read by each speaker in our set, was collected for all the keywords; the phonetically varied training data needed for the filler models for words outside the predefined keyword vocabulary was obtained by collecting 50 read sentences from the standard TIMIT corpus (Lamel, Kassel & Seneff 1986) for each speaker.

5.3 Recording and transcription

The acoustic quality of the target recognition system for VMR is defined by the ORL Medusa system, which incorporates a specified desk mounted microphone and custom designed audio preamplifier stage. However experimental speech recognition systems frequently rely on very high-quality and low-noise acoustic channels. To study the effects of channel quality, VMR1 was recorded using both head and desk microphones in parallel. All data was recorded in a partially soundproof room, so that the basic recorded signals would be as noiseless as possible. The effect of various noise conditions likely to encountered in the real Medusa environment will be investigated later by mixing separately recorded noise with the recorded speech signal.

The prompted messages were fully transcribed by hand, including not only non-speech events, for example loud breaths and tongue clicks, but also disfluencies such as partially spoken words, pauses and hesitations such as "um" and "ah". Non-speech events were transcribed in accordance with the Wall Street Journal data collection procedures (Garafolo, Paul & Phillips 1993). Basic punctuation was also added for ease of reading. This full transcription was required both for word spotter evaluation purposes and also for retrieval performance testing and bench marking.

5.4 VMR1 database details

In total therefore, VMR1 consisted, for each speaker, of:

1. 77 read sentences ("r" data): sentences containing keywords, constructed such that each keyword occurred a minimum of five times.
2. 170 isolated keywords ("i" data): 5 occurrences of each of the 35 keywords spoken in isolation.
3. 150 read sentences ("z" data): phonetically-rich sentences from the TIMIT corpus.
4. 20 natural speech messages ("p" data): the response to 20 unique prompts from 4 categories.
5. 20 "tags" ("t" data): natural speech responses to a prompt requesting a summary for each of their "p" messages.

The "r," "i," and "z" sets are for use as training data; the "p" and "t" material, along with their transcriptions, serves as a test corpus for both keyword spotting and preliminary IR experiments (though the tag data was not used for the work reported here). The messages, the prime material for actual retrieval, average about 7 keywords (tokens) each; they vary somewhat in length, but a two-minute upper limit was imposed in recording. The amount of training and test data we have (5 hours each) is typical of

other speech recognition development databases, though very small in terms of retrieval testing.

6 Word searching

In acoustic processing, natural language words may be treated as integral wholes or as composed from constituent units like syllables or phones; and as noted, hidden Markov modelling can in principle be used for any kind of unit in context. However while integral words may be appropriate for some applications, a compositional approach using underlying phones is more flexible and refined, and has therefore been used in all our experiments. Whole words are thus characterised as strings of phones drawn from the set of about 50 phones representative of spoken English; and words are identified, and distinguished from preceding or following other words, noise, or silence, by Markov modelling. The likelihood that a file word is the required one is the word occurrence's acoustic score, to which a threshold may be applied to remove the majority of false alarms.

For experimental purposes the message file has been preprocessed to obtain an inverted keyword file. This preprocessing is in fact done by the word spotting techniques described below, and there are therefore in our experimental system several inverted files, each representing a different a-score threshold. Word spotting thus plays a critical role in forming the retrieval search files, as well as being the test area for spoken word recognition techniques. With large message files it would probably be necessary to invert explicitly on variant forms of words, but we judged that the full-word models for our keywords would subsume the variants actually occurring in the files. This judgement was shown to be sound since examination of the acoustic stems of keywords occurring in VMR1 revealed the vast majority to be identical to the fixed keyword form used by the word spotter. A preprocessing strategy like that used for our Stage 1 tests is however only viable with a fixed keyword vocabulary: we return to the implications of using open search terms in Section 8.

6.1 Word spotting experiments

Our initial experiments were designed just to study keyword spotting from the acoustic point of view, identifying individual keyword occurrences in the message set. For this we used the HMM toolkit, HTK, developed in Cambridge (Young et al. 1993), which applies fast algorithms both in training HMM parameters and in finding the most likely word model sequence given unknown speech (for full details see Jones et al. (1995)).

After processing the raw acoustic data to obtain a spectral representation, we constructed compositional keyword models [2] and monophone filler models for each speaker; we used the "i" and "r" data for the former, and the TIMIT sentence "z" data along with the non-keyword parts of the "r " data for the latter. Hence each phone within each individual keyword was trained uniquely within its acoustic context from the 10 read examples of the keyword. We have separate model sets for the two microphones.

Keyword spotting is done with a two-pass recognition procedure using Viterbi decoding, in a manner similar to Rose and Paul's (Rose 1991) The first step analyses the stored speech just for filler monophones, using a filler network. The second identifies keywords

[2] Via a pronouning dictionary.

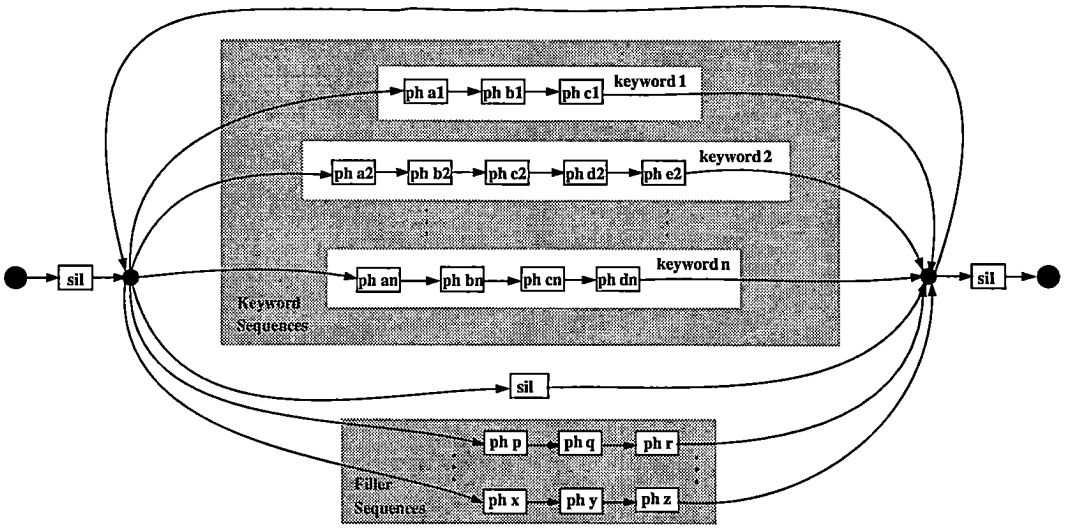

Figure 1: Keyword recognition network.

in parallel against a background of noise and silence using a network combining individual keyword nets, the filler net and silence, as illustrated in Figure 1. The recognition output in the former case is a sequence of filler phones and silence, and in the latter case a sequence of putative keywords, filler phones and silence. The acoustic score for each putative keyword occurrence found in the latter pass is computed by calculating the ratio between the Viterbi score for the keyword and the Viterbi score for the same section of speech data in the first pass. This procedure is described in more detail in (Knill & Young 1994). In practice we found it necessary to compensate for limited training data by using phone triple filler models to discourage the word spotter from fitting the filler model to actual speech occurrences in the recognition pass that includes the keyword models.

The accepted figure-of-merit (FOM) for word spotting is defined as the average percentage of correctly detected keywords as the a-score threshold is varied from one to ten false alarms per keyword per hour. The keyword spotting results were evaluated against time-aligned text transcriptions containing the keywords. The FOMs for the two microphones, averaged across both the 15 talkers and 35 keywords, are 81.2% for head and 76.4% for desk. The receiver operating characteristic (ROC) curve for the two microphones, plotting percent correct against thresholds set to range from 1 to 10 false alarms, is given in Figure 2. This shows that the desk microphone (the realistic office case) is not much inferior to the head one, and also that a high performance level can be obtained for only a few false alarms. These results were exploited for our message retrieval experiments.

7 Message retrieval experiments

Our tests so far have used VMR1 with two different request sets defining two retrieval test collections, VMR Collection 1a and VMR Collection 1b. The primary purpose of these tests has been to establish that spoken document retrieval is feasible and viable. For this purpose we have calibrated against text matching performance. However the tests have also, equally importantly, been used to experiment with different term weighting

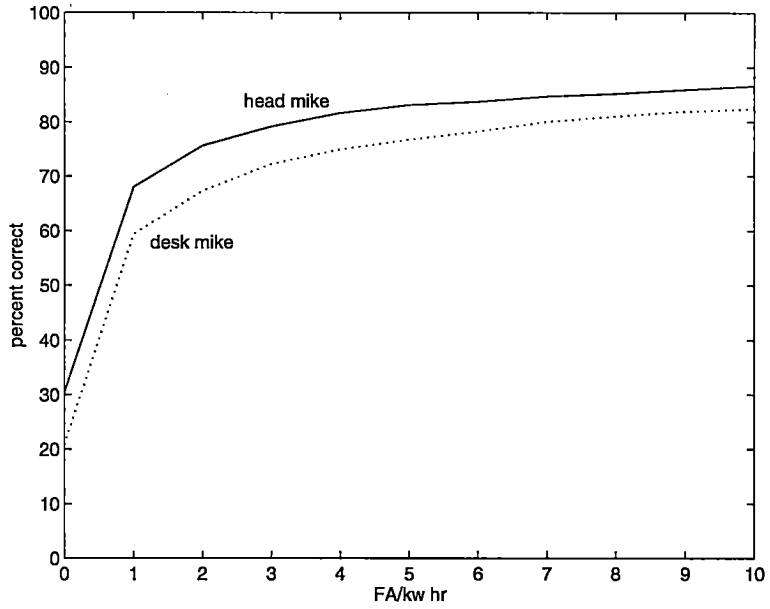

Figure 2: VMR Task Keyword Spotting ROC

schemes (cf Robertson & Sparck Jones (1994), Robertson et al. (1995)), to see whether using spoken as opposed to written documents affects these schemes and in particular their comparative performance. We have naturally also used these tests to check the relative performance for head and desk microphones implied by the word-spotting experiments, and we have used them to study the impact of training data differences on recogniser performance: this is important in relation to speaker-independent systems.

From the point of view even of past retrieval testing, let alone present-day TREC, a file of 300 documents is extremely small. But the constraints imposed by the need to calibrate speech retrieval performance by reference to transcribed data are quite severe, and we thought it important to make some initial trials to see whether actual rather than simulated retrieval for naturally-spoken material was practicable. Of course one consequence of such small data sets, quite apart from other factors mentioned below, is that absolute performance levels have no real meaning, and performance differences have to be treated with caution. The question of larger data sets is considered later.

Note that, as mentioned earlier, in these Stage 1 tests the search queries are confined to the fixed keyword set: we anticipate improved performance and flexibility when we allow open search terms.

7.1 Methodology

7.1.1 Collection 1a

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. To reduce variations in word form, query words were suffix stripped to stems using the standard Porter algorithm (Porter 1980). Queries were formed from the prompts by selecting those stems also found in a stemmed list of the keywords. For example, given the prompt

```
Your current project is lagging behind schedule. Send a message
pointing this out to the other project management staff. Suggest
some days and times over the next week when you would be willing
to hold a meeting to discuss the situation.
```

the following query was obtained:

```
project messag project manag staff time meet.
```

With this expedient we obtained a total set of 50 requests with corresponding simple term list queries, averaging 4.6 distinct terms (average 5.68 occurrences).

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 keywords, 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.

As mentioned earlier, we are only working with written queries in this project. Moreover as the messages have been preprocessed, identifying search term stem occurrences in the files is a purely mechanical operation via the written forms of the keywords that act as access points to the inverted file.

7.1.2 Collection 1b

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. As for the data collection, the category subsets were ordered using a Latin square to prevent sequencing effects. 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. Subjects were also shown the list of otherwords associated with the category.

Relevance Assessment Each request was converted to a query consisting only of the suffix stripped fixed keywords present in the request. For example, given the request,

```
In what ways can the windows interface of a workstation be
personalised.
```

the following query was obtained:

```
window interfac workstat
```

The query was used to score each document transcription in the message archive; keyword terms in the messages were collection frequency weighted (Robertson & Sparck Jones 1994). The messages were then ranked in order to decreasing 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 Collection 1b 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 there were 35 messages to assess for each request. The order of these messages was randomised using a random number generator seeded with a unique request number.

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, Foote & Sparck Jones 1995).

Apart from the greater realism, the main differences between Collection 1a and Collection 1b were that there were far fewer terms per query for the latter, an average of 2.58 distinct terms (average 2.64 occurrences, but one case 0), while 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 17.22 highly or partially relevant.

7.2 Calibration via text retrieval

Retrieval performance for speech documents can be expected to suffer degradation relative to that for text documents due to either misses or false alarms. The extent of the degradation can be measured, when transcribed texts are available, by comparing performance for spoken word spotting results with that for the transcriptions. We used our transcribed corpus to provide us with this performance standard.

However as indicated earlier with the 'Hello Kate' example, there is a further problem with word spotting in that unrelated acoustic events will often resemble valid keywords: even the most accurate acoustic models cannot discriminate in these cases. The output of an ideal word spotter that reports all keyword phone sequences thus provides a more legitimate standard of comparison than text. We simulated this ideal 'phonetic text' performance by scanning phonetic transcriptions of the messages for phone sequences that match those of a keyword. In the tables which follow the two forms of reference are labelled *text* and *phonetic* respectively.

7.3 Spoken message retrieval performance

Our tests have compared unweighted term (*uw*) matching performance with two forms of weighting. As mentioned, the aim was primarily to check that the same sort of comparative performance was obtained as for text (insofar as this could be done with a small collection). We thus used the conventional collection frequency weight (*cfw*) (alias inverse document frequency weights), and the 'combined weight' (*cw*), incorporating within-document term frequencies, normalised for document length, defined in (Robertson & Sparck Jones 1994)

and derived in (Robertson & 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,

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

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. The combined weight constant K has to be tuned empirically: after testing we set $K = 1$.

To measure performance we used only precision metrics, namely precision at selected document cutoff levels in the ranked search output, and average precision as defined in the TREC evaluation (Harman 1994). We believe that document cutoff is highly appropriate where it is very likely that the user will not be willing to consider many output documents, in this case because listening to them takes time. The implementation for average precision covers a ranking of the entire collection.

The word spotter outputs a list of putative keyword hits and associated acoustic scores. As these are probabilistic, it would naturally appear sensible to combine information they provide with that given by the term weighting schemes. However it is not completely clear how this should be done (though we are currently investigating the question), and as the basic retrieval schemes require only the presence/absence of a keyword in a message, we have applied a threshold to the a-scores. This has in any case the advantage that, because acoustic false alarms typically score worse than true hits, thresholding removes a greater proportion of the false alarms. Clearly, it is desirable in practice to choose an operating threshold that optimises retrieval performance in trading false alarms against 'pseudo'-misses, ie hits with scores below the threshold. This question is discussed in more detail in Jones et al. (1995): for the Stage 1 studies the optimal threshold values were chosen *a posteriori* for both head and desk microphone systems, though an *a priori* fixed threshold value would have to be used in an operational system.

We present our retrieval results first for Collection 1a (Tables 1, 2, 3), and then for Collection 1b (Tables 4, 5, 6, 7, 8, 9). As we have tried the same weighting schemes for both collections the consequences of collection differences are easily seen in the parallel tables; however for Collection 1b there are additional figures for the full as opposed to highly relevant assessments. We comment on the results for each collection individually, and then discuss the complete set of tests.

7.3.1 Collection 1a results

Table 1 shows retrieval performance for the standard transcribed messages (*text*) and for the ideal phonetic text (*phonetic*), using different cutoff matching scores. The behaviour is as expected: first, the text transcription performs better than the phonetic reference; second, when weighting is introduced cfw weighting gives a substantial improvement in performance over the unweighted case, and cw in turn does better than cfw. It must however again be emphasised that with these artificial queries and assessments, absolute performance values are not indicative of expected real performance levels; and equally that while our claim is that the relative orderings should hold for the weighting schemes, the small collection means that the observed differences must be treated with caution.

Weight Scheme		Text			Phonetic		
		uw	cfw	cw	uw	cfw	cw
Precision	5 docs	0.264	0.296	0.300	0.248	0.292	0.308
	10 docs	0.222	0.250	0.270	0.216	0.240	0.268
	15 docs	0.192	0.213	0.236	0.189	0.211	0.233
	20 docs	0.170	0.193	0.208	0.169	0.183	0.203
Av Precision		0.293	0.332	0.358	0.279	0.317	0.349

Table 1: Collection 1a: Retrieval precision values for text transcription and ideal phonetic word spotter.

Weight Scheme		Head					
		uw		cfw		cw	
		th_{mco}	th_{map}	th_{mco}	th_{map}	th_{mco}	th_{map}
Precision	5 docs	0.248	0.232	0.288	0.256	0.272	0.260
	10 docs	0.212	0.192	0.218	0.222	0.244	0.234
	15 docs	0.164	0.169	0.185	0.195	0.211	0.213
	20 docs	0.143	0.156	0.160	0.171	0.192	0.187
Av Precision		0.256	0.259	0.293	0.295	0.311	0.316

Weight Scheme		Desk					
		uw		cfw		cw	
		th_{mco}	th_{map}	th_{mco}	th_{map}	th_{mco}	th_{map}
Precision	5 docs	0.252	0.244	0.264	0.260	0.272	0.272
	10 docs	0.176	0.182	0.224	0.214	0.238	0.238
	15 docs	0.164	0.167	0.192	0.191	0.210	0.210
	20 docs	0.141	0.142	0.173	0.166	0.177	0.177
Av Precision		0.232	0.241	0.279	0.283	0.299	0.299

Table 2: Collection 1a: Retrieval performance for head and desk microphone data, thresholds chosen to maximise precision cutoff values (th_{mco}) and average precision (th_{map}).

		Text	Phonetic	Head	Desk
cdf	Relative to Text	100%	95.5%	88.8%	85.2%
	Relative to Phonetic	—	100%	93.2%	89.4%
cw	Relative to Text	100%	97.5%	88.3%	83.5%
	Relative to Phonetic	—	100%	90.5%	85.7%

Table 3: Collection 1a: Relative average precision retrieval performance.

Weight Scheme		Text			Phonetic		
		uw	cfw	cw	uw	cfw	cw
Precision	5 docs	0.342	0.350	0.342	0.338	0.346	0.354
	10 docs	0.281	0.308	0.294	0.288	0.321	0.310
	15 docs	0.260	0.297	0.299	0.263	0.301	0.299
	20 docs	0.242	0.281	0.280	0.251	0.283	0.282
Av Precision		0.296	0.332	0.346	0.302	0.339	0.355

Table 4: Collection 1b: Retrieval precision values for text transcription and ideal phonetic word spotter with highly relevant message set.

Weight Scheme		Head					
		uw		cfw		cw	
		th_{mco}	th_{map}	th_{mco}	th_{map}	th_{mco}	th_{map}
Precision	5 docs	0.333	0.333	0.358	0.350	0.346	0.333
	10 docs	0.265	0.265	0.306	0.321	0.308	0.296
	15 docs	0.238	0.238	0.283	0.285	0.283	0.289
	20 docs	0.207	0.207	0.253	0.260	0.270	0.266
Av Precision		0.265	0.265	0.310	0.312	0.320	0.330

Weight Scheme		Desk					
		uw		cfw		cw	
		th_{mco}	th_{map}	th_{mco}	th_{map}	th_{mco}	th_{map}
Precision	5 docs	0.296	0.296	0.308	0.308	0.338	0.338
	10 docs	0.273	0.273	0.300	0.300	0.302	0.302
	15 docs	0.246	0.246	0.274	0.274	0.282	0.282
	20 docs	0.215	0.215	0.245	0.245	0.254	0.254
Av Precision		0.254	0.254	0.296	0.296	0.307	0.315

Table 5: Collection 1b: Retrieval performance for head and desk microphone data for highly relevant message set, thresholds chosen to maximise precision cutoff values (th_{mco}) and average precision (th_{map}).

		Text	Phonetic	Head	Desk
cdf	Relative to Text	100%	102.4%	94.00%	89.4%
	Relative to Phonetic	—	100%	91.8%	87.3%
cw	Relative to Text	100%	102.7%	95.3%	91.2%
	Relative to Phonetic	—	100%	92.9%	88.9%

Table 6: Collection 1b: Relative average precision retrieval performance for highly relevant message set.

Weight Scheme		Text			Phonetic		
		uw	cfw	cw	uw	cfw	cw
Precision	5 docs	0.472	0.488	0.508	0.472	0.484	0.508
	10 docs	0.430	0.466	0.462	0.434	0.474	0.464
	15 docs	0.391	0.436	0.448	0.396	0.444	0.443
	20 docs	0.371	0.418	0.425	0.375	0.419	0.427
Av Precision		0.403	0.450	0.460	0.409	0.455	0.464

Table 7: Collection 1b: Retrieval precision values for text transcription and ideal phonetic word spotter with full relevant message set.

Weight Scheme		Head					
		uw		cfw		cw	
		th_{mco}	th_{map}	th_{mco}	th_{map}	th_{mco}	th_{map}
Precision	5 docs	0.492	0.492	0.516	0.516	0.492	0.476
	10 docs	0.400	0.400	0.434	0.434	0.456	0.452
	15 docs	0.363	0.363	0.413	0.413	0.429	0.431
	20 docs	0.330	0.329	0.381	0.381	0.408	0.401
Av Precision		0.364	0.364	0.410	0.410	0.418	0.423

Weight Scheme		Desk					
		uw		cfw		cw	
		th_{mco}	th_{map}	th_{mco}	th_{map}	th_{mco}	th_{map}
Precision	5 docs	0.412	0.412	0.444	0.444	0.496	0.496
	10 docs	0.384	0.384	0.430	0.430	0.436	0.436
	15 docs	0.368	0.368	0.399	0.399	0.416	0.416
	20 docs	0.324	0.324	0.370	0.370	0.384	0.384
Av Precision		0.333	0.333	0.385	0.385	0.409	0.409

Table 8: Collection 1b: Retrieval performance for head and desk microphone data full relevant message set, thresholds chosen to maximise precision cutoff values (th_{mco}) and average precision (th_{map}).

		Text	Phonetic	Head	Desk
cdf	Relative to Text	100%	101.3%	91.1%	85.7%
	Relative to Phonetic	—	100%	90.0%	84.6%
cw	Relative to Text	100%	101.0%	92.1%	88.9%
	Relative to Phonetic	—	100%	91.2%	88.0%

Table 9: Collection 1b: Relative average precision retrieval performance, full relevant message set.

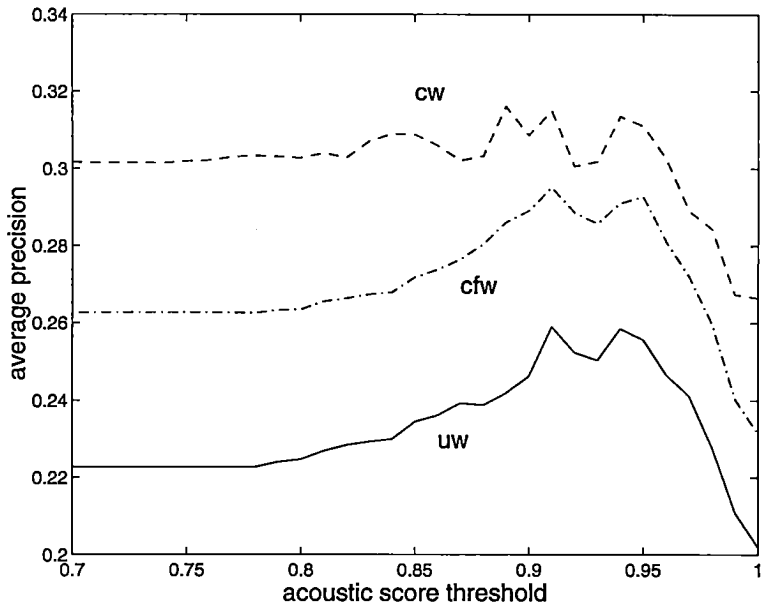

Figure 3: Collection 1a: Average retrieval precision for head microphone data versus threshold.

Retrieval performance results for head and desk microphones, at document cutoffs, are shown in Table 2. The acoustic threshold th_{mco} gives best performance at the 5 document cutoff, and threshold th_{map} best average precision. Again application of cfw and, further, of cw weighting is highly advantageous.

Finally, Table 3 shows average precision performance for acoustic matching using head and desk microphones compared with both transcribed and ideal phonetic text standards. These results show that ideal phonetic retrieval performance is degraded by about 5% relative to that of the standard text transcriptions, because of homophones, so retrieval using even a perfect word spotter will not perform as well as retrieval from text. However, even for an imperfect word spotting system, and considering both head and desk microphones, retrieval performance is encouragingly around 90% of the ideal phonetic figure. As anticipated from the lower FOM reported earlier, retrieval performance for the desk microphone is slightly lower, though it appears it will still be good enough for the eventual Medusa system.

One reason for the good performance of the retrieval system is the inherent robustness of term weighting with respect to false alarms. Cfw weighting penalises both frequently occurring, and hence undiscriminating keywords (as in the text case), and also keywords having high numbers of acoustic false alarms across the document set. Cw moreover, as well as being generally preferable on independent grounds, is yet more robust in relation to speech. Figure 3 shows spoken message retrieval performance for Collection 1a using head microphone data at different acoustic score thresholds for uw, cfw and cw schemes. It can be seen from this figure that the performance trends observed for *a posteriori* best performance thresholds are consistent across the different threshold levels; and also, significantly, that as well as achieving the best retrieval performance in absolute terms, the cw scheme is also less sensitive to the choice of threshold than the other schemes.

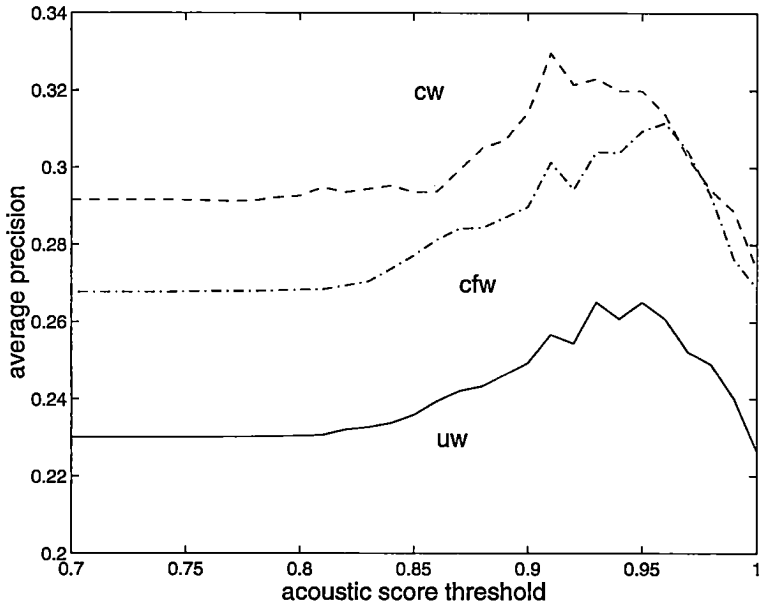

Figure 4: Collections 1b: Average retrieval precision for head microphone data versus threshold for the highly relevant message set.

7.3.2 Collection 1b results

As Collection 1b is rather more realistic than Collection 1a (or at least is less artificial), we replicated all the tests done with Collection 1a on Collection 1b. Tables 4, 5, 6 thus show comparative performance for the three retrieval strategies for the two reference standards, for spoken document retrieval for the head and desk microphones, and speech performance relative to the two standards. These tables give the figures for the highly relevant documents only: performance for the highly and partially relevant documents taken together is given in Tables 7, 8, 9.

The results again show the three strategies in the same rank order, with cw better than cfw better than uw, and results for the head microphone better than the desk one. (These comparisons are subject to the same caveats as those for Collection 1a.) The difference between Collection 1a and Collection 1b is that far the latter performance is better against the phonetic than against the text standard. This is somewhat surprising, but seems to be attributable to the fact that the queries have so few terms. Where there are only one or two query terms, the retrieval of text transcriptions is particularly susceptible to failures arising from occasional inconsistent suffix stripping of different word forms. This problem is less likely in the phonetic case where the matching is based on comparison of the document contents to the acoustic composition of the standard keyword form.

Figure 4 shows the effect of differing thresholds on retrieval performance for Collection 1b for the highly relevant message set.

7.3.3 Discussion of test results

Apart from the differences for the two standards, the picture for the two collections, with their rather different requests and assessments, is similar. This includes good levels of

performance for spoken document retrieval relative to the standards, suggesting that reasonable speech retrieval performance can be envisaged in real situations like the intended VMR office context, even if the circumstances in which our Collection 1 materials were created and recorded were favourable ones.

By comparison with James's tests (James 1995), our absolute levels of performance were lower. This is not surprising since James's data, consisting of well-defined, summary-type documents, provided unnaturally favourable conditions. The categories covered in our message set are closely related and much less orthogonal than the individual news stories used as documents in his database. However we anticipate that use of an open search term vocabulary, as in James's experiments, would lead to an improvement in absolute performance levels in our system. The general trend towards better performance as more factors are taken into account in weighting is the same in both cases, while we also achieve performance of around 90% in comparison with text transcriptions.

Both sets of experiments relied on known speakers, a limitation that has to be removed before operational deployment is possible. The major advances in the experiments described above, compared with James's, relate mainly to the composition and acoustic quality of the spoken documents. The messages in VMR1 are entirely unscripted and spoken by naive talkers. As noted earlier, recognition on this type of spontaneous spoken material is much more difficult than on read speech. Further, our comparative tests between head and desk microphone data show that retrieval performance is only slightly degraded by reduced acoustic channel quality.

8 Current and Future Work

Referring back to our description of the VMR Project structure, we have successfully completed Stage 1, implementing and testing a retrieval system using a fixed keyword vocabulary for a closed user community (i.e. speaker set). We have also implemented a basic prototype system with appropriate user interface which, among more obvious facilities, offers a graphical display showing the location of keywords in the recorded signal along with their 'brightness' i.e. certainty as reflected in their acoustic score: for further details see Brown, Foote, Jones, Sparck Jones & Young (1994). We have not, however, been able to try out the use of header information in hybrid searches, e.g. for filtering messages by date or sender before a term search, since we neither have (nor expect to have) a file with the appropriate properties.

We have, on the other hand, already begun to address two major issues for the next stages of the work. One is the progression to speaker-independent word recognition, and the other is the provision of a significantly larger test collection. In addition, we are already planning for retrieval using an open search vocabulary, as specified in the Project plan and clearly required for practical speech retrieval applications.

8.1 Training data

In the longer term we want to have speaker-independent rather than speaker-dependent speech recognition and also, preferably, domain independent recognition. A significant issue for the development of speaker-independent models is the collection of a suitable acoustic model training corpus. Fortunately, a recent collection of spoken British English

at the Cambridge University Engineering Department provides a suitable corpus. This corpus, called WSJCAM0 (Robinson et al. 1995), consists of sentences taken from the Wall Street Journal spoken by 100 native British English speakers divided equally between male and female speakers and covering a wide age range and differing regional backgrounds. Hence a word spotter trained using this corpus should be able to operate successfully on any spoken British English data.

Tables 10 shows initial results for retrieval experiments using speaker-independent head data models for Collection 1a and 1b requests. These results are only indicative, but although the performance is somewhat degraded relative to the dependent models, this level of performance may be adequate in operation.

We are currently investigating the potential improvements to word spotting performance with these completely speaker-independent models by adapting them to specific speakers (Leggetter & Woodland 1995). A full description of this work is given in (Foote, Jones, Sparck Jones & Young 1995).

8.2 New test data

The need for reference text is a major restriction on speech retrieval experiment, and its provision is a major challenge. An important point is that it is not only required as a standard for calibration; it is effectively mandatory, for big collections, as a base for relevance assessment, since listening to large amounts of recorded material in order to determine document relevance is extremely burdensome. At the same time, it is necessary to increase the size of our test collections. We are currently, since we do not expect any volume of Medusa data to become available in the short term, investigating the use of television news bulletins which have accompanying teletext subtitles. Even if the speech-text relationship is not one-to-one, it may be close enough for the two purposes mentioned. This material is also quite naturally uttered, even if slightly more formally than typical Medusa mail messages would be.

8.3 Open search vocabulary

The fixed (and small) keyword vocabulary we have used so far is manifestly too limiting. Some work has already been done in Cambridge on single phone index keys (James & Young 1994) which can in principle support open vocabulary searching via the spoken discourse analogue of written text position indicators - we may call these 'phone position indicators', *ppi*. Thus words could be identified as phone sequences via the ppis. However while we intend to pursue this idea with Collection 1, there are extremely daunting problems to overcome in any attempt to apply this idea with large files, both because the indexes are very bulky and because each phone has associated uncertainty.

8.4 Conclusion

Other topics to be addressed are the possible use of actual acoustic scores, and development of the user interface. Finally, there are significant problems to tackle, given the lack of prior information about word incidences in the file, in implementing any iterative searching with query expansion.

Weight Scheme		Head					
		uw		cfw		cw	
		th_{mco}	th_{map}	th_{mco}	th_{map}	th_{mco}	th_{map}
Precision	5 docs	0.208	0.208	0.236	0.236	0.276	0.276
	10 docs	0.170	0.170	0.188	0.188	0.226	0.226
	15 docs	0.145	0.145	0.167	0.167	0.197	0.197
	20 docs	0.128	0.128	0.146	0.146	0.164	0.164
Av Precision		0.210	0.210	0.240	0.240	0.290	0.290

Results for Collection 1a.

Weight Scheme		Head					
		uw		cfw		cw	
		th_{mco}	th_{map}	th_{mco}	th_{map}	th_{mco}	th_{map}
Precision	5 docs	0.300	0.300	0.317	0.317	0.321	0.317
	10 docs	0.233	0.233	0.250	0.250	0.269	0.288
	15 docs	0.215	0.215	0.248	0.248	0.247	0.257
	20 docs	0.202	0.202	0.234	0.234	0.229	0.244
Av Precision		0.236	0.236	0.273	0.273	0.262	0.299

Results for Collection 1b, highly relevant message set.

Weight Scheme		Head					
		uw		cfw		cw	
		th_{mco}	th_{map}	th_{mco}	th_{map}	th_{mco}	th_{map}
Precision	5 docs	0.416	0.416	0.432	0.428	0.456	0.444
	10 docs	0.340	0.340	0.364	0.390	0.420	0.410
	15 docs	0.315	0.315	0.357	0.375	0.385	0.388
	20 docs	0.300	0.300	0.343	0.353	0.366	0.371
Av Precision		0.319	0.319	0.358	0.360	0.381	0.384

Results for Collection 1b, full relevant message set.

Table 10: Retrieval performance for speaker independent acoustic models, thresholds chosen to maximise precision cutoff values (th_{mco}) and average precision (th_{map}).

However the results we have obtained so far in the VMR Project suggest that speech recognition and document retrieval technologies have developed to a point were spoken document retrieval is a feasible proposition.

References

Brown, M. G., Foote, J. T., Jones, G. J. F., Sparck Jones, K. & Young, S. J. (1994), Video Mail Retrieval Using Voice: An overview of the Cambridge/Olivetti retrieval system, *in* 'Proceedings of the ACM Multimedia '94 Conference Workshop on Multimedia Database Management Systems', San Francisco, CA, pp. 47–55.

Burrows, M. (1991), DEC Systems Research Centre, Palo Alto, personal communication.

Foote, J. T., Jones, G. J. F., Sparck Jones, K. & Young, S. J. (1995), Talker-independent keyword spotting for information retrieval, *in* 'Proceedings of Eurospeech 95', ESCA.

Garafolo, J., Paul, D. & Phillips, M. (1993), CSR WSJ0 Detailed orthographic transcription (.dot) specification, (ftp: jaguar.ncsl.nist.gov /csr/csr-dot-spec.doc).

Glavitsch, U. & Schäuble, P. (1992), A system for retrieving speech documents, *in* 'Proceedings of SIGIR '92', ACM, pp. 168–176.

Glavitsch, U., Schäuble, P. & Wechsler, M. (1994), 'Metadata for integrating speech documents in a text retrieval system', *SIGMOD Record* **23**(4), pp. 57–63.

Harman, D. K. (Ed.) (1994), *The Second Text REtrieval Conference (TREC-2)*, Special Publication 500-215, National Institute of Standards and Technology, Gaithersburg MD.

James, D. A. & Young, S. J. (1994), A fast lattice-based approach to vocabulary independent wordspotting, *in* 'Proceedings of ICASSP 94', IEEE, pp. I–(377–380).

James, D. A. (1995), *The application of classical information retrieval techniques to spoken documents*, Dissertation, Engineering Department, University of Cambridge.

Jones, G. J. F., Foote, J. T., Sparck Jones, K. & Young, S. J. (1994), Video mail retrieval using voice: Report on keyword definition and data collection, Technical Report 335, Computer Laboratory, University of Cambridge.

Jones, G. J. F., Foote, J. T., Sparck Jones, K. & Young, S. J. (1995), Video mail retrieval: the effect of word spotting accuracy on precision, *in* 'Proceedings of ICASSP 95', Vol. 1, IEEE, pp. 309–312.

Jones, G. J. F., Foote, J. T. & Sparck Jones, K. (1995), Video mail retrieval using voice: Report on collection of naturalistic requests and relevance assessments, VMR Project Working Document.

Knill, K. M., & Young, S. J. (1994), Speaker dependent keyword spotting for hand-held devices, Technical Report 193, Cambridge University Engineering Department, July 1994.

Kupiec, J., Kimber, D. & Balasubramanian, V. (1994), Speech-based retrieval using semantic co-occurrence filtering, *in* 'Proceedings of HLT 94', (ARPA), San Francisco: Morgan Kaufmann, pp. 350–354.

Lamel, L. F., Kassel, H. K. & Seneff, S. (1986), Speech database development: Design and analysis of the acoustic-phonetic corpus, *in* 'Proceedings of the DARPA Speech Recognition Workshop', pp. 26–32.

Leggetter, C. & Woodland, P. (1995), Flexible speaker adaptation using maximum likelihood linear regression, *in* 'Proc. ARPA Spoken Language Technology Workshop', Barton Creek.

McDonough, J., Ng, K., Jeanrenaud, P., Gish, H. & Rohlicek, J. R. (1994), Approaches to topic identification on the switchboard corpus, *in* 'Proceedings of ICASSP 94', IEEE.

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

Rabiner, L. R. (1989), 'A tutorial on hidden Markov models and selected applications in speech recognition', *Proceedings of the IEEE* **77**(2), 257–286.

Robertson, S. E. & Sparck Jones, K. (1994), Simple, proven approaches to text retrieval, Technical Report 356, Computer Laboratory, University of Cambridge.

Robertson, S. E. & Walker, S. (1994), Some simple effective approximations to the 2–Poisson model for probabilistic weighted retrieval, *in* 'Proceedings SIGIR '94', ACM, Dublin, pp. 232–241.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M. & Gatford, M. (1995), Okapi at TREC-3, *in* D. K. Harman, ed., 'The Third Text REtrieval Conference (TREC-3)'. in press.

Robinson, T., Fransen, J., Pye, D., Foote, J., & Renals, S. (1995) WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition, *in* 'Proceedings of ICASSP 95', Vol. 1, IEEE, pp 81–84.

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

Schäuble, P. & Glavitsch, U. (1994), Assessing the retrieval effectiveness of a speech retrieval system by simulating recognition errors, *in* 'Proceedings HLT 94', ARPA, pp. 347–349.

Tague, J. M. (1981), The pragmatics of information retrieval experimentation, *in* K. Sparck Jones, ed., *Information Retrieval Experiment*, London: Butterworths, chapter 5, pp. 59–102.

Wilcox, L. D. & Bush, M. A. (1992), Training and search algorithms for an interactive wordspotting system, *in* 'Proceedings of ICASSP 92', Vol. II, IEEE, pp. 97–100.

Woodland, P. C., Leggetter, C. J., Odell, J. J., Valchev, V. & Young, S. J. (1995), The 1994 HTK large vocabulary speech recognition system, *in* 'Proceedings of ICASSP 95', Vol. 1, IEEE, pp. 73–76.

Young, S. J., Woodland, P. C. & 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.

Young, S. J., Woodland, P. C. & Byrne, W. J. (1994) 'Spontaneous speech recognition for the Credit Card corpus using the HTK Toolkit', *IEEE Transactions on Speech and Audio Processing*, **2**(4), 615-621.