ELECTRONIC WORKSHOPS IN COMPUTING

Series edited by Professor C.J. van Rijsbergen

Ian Ruthven (Ed)

Miro'95

Proceedings of the Final Workshop on Multimedia Information Retrieval
(Miro '95)
Glasgow, Scotland
18-20 September 1995

Paper:

Video Mail Retrieval Using Voice: An Overview of the Stage2 System

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

Published in collaboration with the
British Computer Society

©Copyright in this paper belongs to the author(s)

Video Mail Retrieval using Voice: An Overview of the Stage 2 System

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

[†]Computer Laboratory, University of Cambridge
Cambridge, CB2 3QG, U.K.

[‡]Engineering Department, University of Cambridge
Cambridge, CB2 1PZ, U.K.

Abstract

This paper outlines the Video Mail Retrieval (VMR) project at Cambridge University. The goal of the VMR project is to develop 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 by combining state-of-the-art speech recognition with established document retrieval technologies to provide an effective and efficient retrieval tool. Experimental results with a small spoken message collection show that retrieval precision is somewhat dependent on the generality of the acoustic modelling used. For talker-dependent acoustic modelling retrieval performance is around 95% of that observed when text transcriptions of the same files are used. However, even with incorporation of completely open-user talker-independent acoustic models, retrieval performance of about 75% of text can be obtained.

1 Introduction

This paper discusses work to date on the Video Mail Retrieval (VMR) project at Cambridge University. Our objective is to develop a novel multimedia application for the retrieval of spoken documents. The 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 outlines the problems involved, specific strategies being deployed to overcome these, the current system implementation, and the design and results of our retrieval tests to date. We demonstrate that the straightforward probabilistic methods established for text retrieval can be naturally extended to the speech domain; and also that current speech recognition technology can support good message retrieval performance. Here we present results to support these claims and summarise our work to the end of Stage 2 of the VMR project.

Section 2 presents background details of the VMR project: 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 describes the specific objectives of the VMR project, section 4 outlines the project strategy, and section 5 considers problems encountered in spoken document retrieval. Details of our experimental investigations to date are given in section 6. Finally, section 7 comments on our work to date and summarises our planned future research. Much of the material in this paper brings together key results from work previously appearing in [1], [2], [3], and [4].

2 Background

Recent years have seen a rapid expansion in the availability of multimedia applications, including video conferencing, and video and audio mail. Using these systems can create large archives of material which can

pose significant problems since the data is expensive to store and unwieldy to access. A particular problem is that users are unable to find particular stored documents since, unlike text, there is no simple content-based way to search for an individual reference. Manually searching of an archive by listening is significantly more time consuming than a similar search of a text archive since audio browsing is much less efficient than visual browsing.

Little work has been done in the area of spoken document retrieval, and what has been done has been limited in scope or evaluation. Schäuble and others [5] have proposed a system for spoken document retrieval based on predefined acoustic units, and have considered the effect of term occurrence errors of semantic and acoustic origin, but only by simulation. Their most recent work has included some real speech document retrieval results [6]. However, there are significant differences between this work and that reported here and hence it cannot be compared directly.

The VMR project is addressing these problems in multimedia retrieval and browsing by developing a system to retrieve stored video mail messages using voice indexing. A specific goal of the project is to develop a useful retrieval application for the Medusa multimedia environment installed at Olivetti Research Ltd in Cambridge [7]. The following section describes the salient features of the Medusa system.

2.1 The ORL Medusa system

The Medusa Project at ORL is an experimental multimedia system based on multiple streams of digital audio and video sent over a 100 Megabit-per-second switched ATM network [8, 9]. The ATM fabric covers all rooms at the laboratory and some 200 network connections are in place. An ATM network has useful characteristics for multimedia transport: the bandwidth is relatively high, the latency is low, and transit time jitter is also low.

Figure 1: Connection of Medusa Direct Peripherals to an ATM network.

Multimedia input and output functionality has traditionally been provided by cards plugged into a workstation bus. In contrast, Medusa adopts a more flexible approach of making microphones, speakers, cameras and storage systems into first-class network objects. The Medusa hardware design is modular: various ATM direct peripherals are made by adding specialised options cards to an ATMos card. The ATMos card has an ATM network interface, an ARM processor and up to 32 MBytes of memory. A range of Smart ATM Modules (SAMs) have been developed. These are lunchbox-sized units which plug directly into the ATM network. Figure 1 shows connection of various Medusa ATM Direct Peripherals to the network. Figure 2 shows a typical Medusa installation.

An Audio SAM supports a set of ATM microphones and speakers. Audio input signals are sampled at

Figure 2: The Medusa multimedia system.

32kHz using 16-bit coding, although higher sampling rates are possible. Stereo audio can be transmitted across the network using two audio channels enabling near CD quality digital audio playback. The Video SAM is an ATM networked video unit to which one or two colour camera heads can be connected. Colour video is output to either ATM networked workstations or a colour LCD VideoTile as shown to the right of the workstation in figure 2.

Medusa application programs (including the VMR user interface described in Section 6.5) are written in Tcl/Tk [10], a scripting language with useful built-in X widgets. Tcl has been extended at ORL to include commands for initialising Medusa modules; making connections between them; and reading and writing to the controlling attributes on the modules. In the case of an audio source module the attributes include the sampling rate, the number of sample bits, and the quantisation method.

Several large Tcl applications have been written to demonstrate the multi-stream capabilities of Medusa. The Medusa phone application can show small images from up to four cameras from each installation at each end of the conversation, as well as large images from each end which can be either selected manually or automatically. Agent software can use the output from analyser modules attached to the audio and video sources, to automatically select video and audio streams. The Medusa mail application (Figure 3) makes a recording of all four streams of video available from a Medusa system as well as the most appropriate microphone stream. These five media streams are passed through timestamping pipeline modules and recorded onto ATM networked Disc Bricks. During playback, all four recorded views are seen, and one of them can be selected to be shown in large format, as shown in Figure 3.

3 VMR Project Goals

The VMR project goals should be evident from the preceding sections. The primary research issue is how to integrate text retrieval methods with automated speech recognition technology. However, the final objective is to develop a practical spoken document retrieval system for the ORL Medusa environment. A usable tool must rapidly provide robust high retrieval performance using a practical amount of computation and storage. Because speech recognition is computationally expensive, the only practical way to achieve this is to index spoken documents at the time they are added to the archive. To retrieve documents, the pre-computed indexes can then be rapidly searched to find potentially relevant documents. Identifying potentially relevant documents is only part of the solution; the user must be able to select and play back any desired document. Since audio and video consume orders of magnitude more storage than text, this is a non-trivial problem in itself. Further, it is inefficient to play back entire documents when just a small portion is of interest. A truly useful retrieval application should give the user the ability to identify and play back potentially interesting portions of individual messages. Finally, a useful application should be well-integrated into the environment in which it is to be used, in our case Medusa.

Figure 3: Medusa video mail application.

4 The Project Structure

The VMR project has three phases that encompass progressively less restricted and more realistic conditions for spoken document retrieval. In Stages 1 and 2 (the work reported here) searching depends only on a fixed keyword vocabulary, known in advance of search time. Stage 1 assumed a closed talker community responsible for all messages. Speech recognition was restricted to locating instances of the fixed keyword set in acoustically clean speech from a known talker from a set of 15 talkers. In addition, all talkers provided sufficient examples of their speech so that specialised models could be built. Though this is clearly unrealistic in the long run, it did provide a benchmark for performance in later, less favourable conditions.

Stage 2 relaxed the requirement that the talker's identity be known in advance. In addition, more sophisticated retrieval strategies were implemented and additional evaluation material used.

Key results from Stages 1 and 2 are described later in this paper. The third stage, where research is ongoing, extends the search term vocabulary from the small set of known keywords to a potentially open search term vocabulary. An integral part of each stage is the development of the general user interface and ultimately the integration into Medusa.

5 Problems in Spoken Document Retrieval

5.1 Speech Document Indexing Issues

Attempts to retrieve spoken documents encounter similar problems to those associated with text document retrieval but there are further important issues which must be considered. Most obvious among these is that the contents of spoken documents are unknown and hence the initial phase must perform an indexing operation using speech recognition. This speech recognition phase may be carried out in one of two basic ways. Either the speech recognition system may attempt to perform a full transcription of the contents of the documents using a large vocabulary recogniser or recognition may be restricted to a limited and, hopefully, useful set of indexing terms (or *keywords*) selected *a priori*. The recognition used in the Stages 1 and 2 of the VMR project is of the second type. In both of these approaches the indexing vocabulary is limited to that of the recogniser. Large vocabulary systems are now available with vocabularies approaching 100,000 words but this should be compared to the 500,000 word vocabularies encountered in text retrieval systems. Many words, particularly proper nouns, cannot be recognised correctly by either recognition system since they are outside the domain of its vocabulary. This creates a significant search problem which does not exist in text based systems where new document terms are merely added into the inverted file structure.

Additionally, speech recognition is inherently not completely reliable. Even the very best systems will make recognition errors often arising from variable pronunciation or events outside its domain. The recogniser typically maps out-of-vocabulary words to something in its existing vocabulary, which will inevitably result in a recognition error. Short words are more susceptible to recognition errors than longer ones, both because of their inherent greater confusability and the greater tendency to poor articulation. In any case it is important to realise that good text search terms may not be as useful in the speech domain because of their acoustic properties.

5.2 Comparison with Text IR

There are some similarities between term identification for text and spoken document retrieval. For example, in the text case there may be *false alarms* on search terms with multiple senses. If one of these terms is present in a query it will match any occurrence in a document regardless of the sense used in each case. This has been shown to have minimal effect on retrieval effectiveness except for very short queries [11], but is nevertheless a real issue. Also, there may be *misses* on search terms which occur in the documents as synonyms of a query term; although of course there are techniques designed to overcome this problem. Finally there may be query-document term matching errors arising from spelling errors in the documents or query, or inappropriate term stemming.

These problems are also potentially present in spoken document retrieval. However, there are two additional sources of potential error similar in effect to those just described. Acoustic false alarms which occur when the speech recogniser hypothesises the presence of a term when none is actually present, and acoustic misses where the occurrence of a term is not detected by the recogniser.

All these sources of search error may be offset by adding more search terms to the query.

6 Experimental Investigations

This section outlines experimental strategies on the VMR project. The following subsections describe the experimental message archive, acoustic training data for the speech recognition component, indexing of spoken documents via word spotting, retrieval testing and our prototype video mail retrieval application.

6.1 Message Archive

A particular problem which we encountered was the lack of real video mail data for experimental use. Thus we had to engage in a serious collection construction exercise for our initial retrieval test data. Our first message set, VMR1, was designed to satisfy requirements of both document retrieval and the speech recognition. From the document retrieval perspective the database had to consist of messages with the same general properties as could be expected in real video mail messages. 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. At the same time the corpus should have the sort of message similarities and differences that pose challenges for recall and precision typical of expected VMR situations. Messages should also have similar acoustic properties and speaking styles to those found in an operational system, and be of comparable length.

A key issue for system assessment is to evaluate the performance of the speech recognition component and to investigate the extent to which word recognition accuracy affects retrieval performance. For this reason all messages were orthographically transcribed, including marking of pauses, disfluencies, and extraneous noises. This detailed transcription can only be done manually and is very expensive to carry out. For this reason the VMR1 archive is a very small collection from the retrieval point of view. In order for this small database to be viable for retrieval research it had to be carefully structured.

6.1.1 The VMR1 Message Corpus

The structure of the VMR1 archive was derived as follows. Messages were sought on *topics* within a set of *topic categories*. Associated with each category was a set of *keywords* drawn from a small fixed keyword *vocabulary* from which all search terms used in Stages 1 and 2 of the project must be taken. In addition, since the keyword vocabulary is not very large, a set of *other-words* were provided for each category as further prompting and potential search vocabulary. The messages were prompted by using *scenarios* which stimulated the talker to talk on a topic within a category without constraining them to produce messages strictly tied to pre-specified topics. The prompt for each spontaneous message consisted of the scenario and the keywords and other-words for the category. Talkers were asked to favour the use of the listed keywords and other-words, 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 talkers were not shown a complete list of the keywords available, but only those relevant to the current category. However, they were free to use any keyword in any message. The collected messages varied in their individual topics but are clustered round the prompting categories.

The total keyword vocabulary was 35 words, along with a total of 31 related other-words. The keywords were selected manually and contained a mixture of longer more easily recognised words and shorter monosyllabic words. The full list of 35 keywords used was:

active assess badge camera date display document find indigo interface
keyword locate location mail manage meeting message microphone network
output Pandora plan project rank retrieve score search sensor spotting
staff time video windows word workstation

A total of 10 topic categories were defined and a keyword subset associated with each one. The categories were chosen to reflect the anticipated messages of a particular user community, the staff associated with the VMR project. The 10 categories were:

spotting document output retrieval windows
management badge Pandora schedule equipment

Keywords were assigned manually to the categories as being representative of the topics defined within the category. Eg for the category *schedule* the following assignment was made:

schedule -> manage project meeting plan

5 message prompt scenarios were generated for each category. 15 evenly distributed sets of 4 categories were formed. Each category group was assigned to a knowledgeable talker. The talker then recorded a message in response to each prompt, giving a total of 20 messages for each talker, and a total of 300 in the VMR1 archive.

The average message length was approximately 1 minute, giving a total message archive of around 5 hours. The average number of fixed keywords per message was about 7. A detailed description of VMR1 is contained in [12].

6.2 Speech Training Data

Speech recognition systems require acoustic data for the training of recognition models. In Stages 1 and 2 of the VMR project training data was required for fixed keyword models, background models for non-keyword speech, and a model for silence. In Stage 1 individual models were needed for each talker, and so separate training data had to be collected for each model set. In Stage 2 general acoustic recognition models were trained to recognise any talker¹. This requires a large set of training data collected from many different talkers.

6.2.1 Stage 1 Training Data

Each talker provided the following speech training data:

- 77 read sentences (“r” data): sentences containing keywords, constructed such that each keyword occurred a minimum of five times.
- 170 isolated keywords (“i” data): 5 occurrences of each of the 35 keywords spoken in isolation.
- 150 read sentences (“z” data): phonetically-rich sentences from the TIMIT corpus [13].

There were a total of about 5 hours of spoken training data collected from the same 15 talkers who generated the experimental message set.

6.2.2 Stage 2 Training Data

For this stage the WSJCAM0 British English spoken corpus was used. This consists of spoken sentences taken from the Wall Street Journal. 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 [14].

6.2.3 Recording Environment

All speech data was recorded in parallel at 16kHz using both the desk microphone from the Medusa system and a Sennheiser HMD 414 close-talking microphone (as used in many current speech recognition systems). The former represents the operational system and latter functions as an experimental control. The speech recordings were made in a quiet environment.

6.3 Word Spotting

Automatically detecting fixed keywords in unconstrained speech is termed “word spotting” [15]; this technology is the basis of the speech recognition in Stages 1 and 2 of the VMR project. The best-performing word spotters are based on hidden Markov model (HMM) methods, used in successful continuous-speech recognition [16]. A hidden Markov model is a state-based statistical representation of a speech event, typically a word or subword. Different states model differing characteristic speech sounds. A typical subword unit is the phone (sometimes referred to as a phoneme). All words are built from a phone sequence drawn from the set of around 45 distinct phones. Phones vary somewhat with context, i.e. the phones which precede and succeed them. When sufficient training data is available improved recognition can be achieved by modelling this variation.

There exist efficient algorithms for both training HMM parameters and finding the most likely model sequence given unknown speech input. The HTK tool set developed at Cambridge University ([17]) is a powerful and flexible set of software tools for developing HMM applications such as the keyword spotting system presented here.

Figure 4: Keyword recognition network.

6.3.1 Word Modelling

Stage 1 models Whole-word talker-dependent keywords models and phone filler models were constructed for each of the 15 talkers, using the HTK tool set. Each keyword model was trained with the 10 occurrences of the word in the training data. Filler phones models to represent non-keyword speech were trained on the remaining data. The Stage 1 keyword recognition network is shown in figure 4. The filler models here were monophones, phone models which are independent of context. All examples of these phones occurring in the filler training data are used to train a single HMM model of each phone.

Figure 5: Keyword recognition network.

Stage 2 models A set of talker-independent keyword models were formed using using the WSJCAM0 data [14]. The keyword models were built using word-internal triphone HMMs. Word-internal triphones model phone context within words, but do not take into account phonetic variation arising from interaction with the previous or following word. These were generalised using a tree-based clustering technique [18]. This training method enables all possible triphones, biphones and monophones to be modelled. Given such a model set, a particular keyword may be easily modelled by concatenating the appropriate sequence of subword models (obtained from a phonetic dictionary). Biphones are used at the beginning and end of the keyword, while triphones model the internal structure. For example, the keyword “find” is represented by the model sequence $f+ay$ $f-ay+n$ $ay-n+d$ $n-d$. Non-keyword speech is modelled by an unconstrained network of monophones. The Stage 2 keyword recognition network is shown in figure 5.

¹The talker-independent system developed here was trained primarily for native British English talkers.

6.3.2 Keyword Recognition

The speech recognition is performed using the Viterbi algorithm, a standard technique for HMM based speech recognition [16]. The Viterbi algorithm combines the HMM model parameters and spoken data to calculate the most likely state sequence of the HMM models in the recognition vocabulary. The output is the corresponding HMM model sequence.

Word spotting is done with a two-pass recognition procedure [15]. First, the Viterbi decoding is performed on a network of just the filler models, yielding a time-aligned sequence of the maximum-likelihood filler monophones and their associated log-likelihood scores. Secondly, another Viterbi decoding pass is done using the appropriate full network (as shown in figures 4 and 5). Putative keyword hits are rescored by normalising each hypothesis score by the average filler score over the keyword interval. This procedure helps take into account variation in hypothesis score arising from changes in speaking style or background conditions.

In the Stage 1 system, it was necessary to tune the filler models so that they did not match an undue number of keywords. This problem arose because of the limited training data available for the whole-word keyword models. A satisfactory solution was to introduce filler models of common 3-phone sequences (not contained within any of the keywords) by concatenating three monophone models, and adjusting the word transition penalty to penalise the filler sequences (which must be traversed in groups of three).

A similar problem was observed in the Stage 2 system; however, in this case there were too many false alarms. A solution was to introduce a separate transition penalty to the keyword models as shown in figure 5. It is observed that increasing this penalty dramatically reduces the number of false alarms, while only mildly impacting the number of correctly identified keywords. The net effect is similar to, but better than, increasing a cut-off threshold on keyword scores, such that those with low scores are ignored.

6.3.3 Recognition Results

	Data set	
	head (%)	desk (%)
Dependent	81.2	76.4
Independent	69.9	55.9

Table 1: Average Figures of Merit

An accepted figure-of-merit (FOM) for word spotting is defined as the average percentage of correctly detected keywords as a threshold on the putative keyword scores is varied from one to ten false alarms per keyword per hour. The keyword spotting output was scored by comparing it against time aligned manual text transcriptions of the documents. A putative hit is counted as a hit if it overlaps more than half of an occurrence of this keyword in the text transcription. Exact time aligned boundaries of the transcriptions and word spotting output are unlikely to be the same. Word boundaries in the text transcription are determined manually by the transcriber and those from the word spotter are calculated stochastically in the Viterbi decoder. In the speech recognition the Viterbi decoder must form the optimal state sequence for the acoustic data using only available models. Inevitably this optimal fit will slightly distort the word boundaries. The FOM for model sets in the Stage 1 and 2 word spotting systems are shown in table 1, averaged across both the 15 talkers and the 35 keywords. FOMs for the talker-independent models are taken at the best experimentally-determined transition penalty value.

6.3.4 Talker Adaptation

VMR1 is realistic in that it contains talkers with non-British accents. For example, one of our talkers is a native American. This is problematic when using models trained exclusively on British English talkers since the model parameters will not well represent the acoustic content of the speech of these talkers. In an attempt to ameliorate this problem, and increase word spotting performance in general, talker adaptation was investigated. In this procedure a small amount of “adaptation” data is used to generate a modified HMM model set which better represents the speech of the individual talker. The approach chosen was maximum-likelihoods linear regression because it has been shown to improve recognition with a comparatively small

	Data set	
	head (%)	desk (%)
R13	77.1	57.8
R75	80.5	65.5

Table 2: Average Figures of Merit with Talker Adaptation

amount of adaptation data [19]. This method involves adapting only some of the model parameters (the means of the HMM Gaussian mixtures) to increase the likelihood of the adaptation data given the models. Varying amounts of the VMR Stage 1 training corpus were used as enrolment data for talker-adaptation experiments.

Word spotting performance using talker adaptation is shown in table 2. The *R13* row used 13 utterances of enrolment data containing in all 2 occurrences of each keyword. The *R75* row used the full 75 “r” sentences from the Stage 1 training material, containing 5 utterances of each keyword. Adaptation does not uniformly improve performance for all talkers. However the average increase is substantial, and is particularly dramatic for our American English talker. As shown in table 2 using a small amount of enrolment data improved the FOM performance substantially.

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

Further details of our word spotting systems and corresponding experimental results are contained in [2, 3].

6.4 Message Retrieval

6.4.1 Requests and relevance assessments

Our retrieval tests so far have used VMR1 with two different request sets defining two retrieval test collections, VMR1a and VMR1b. The primary purpose of these tests has been to establish that spoken document retrieval is feasible and viable.

VMR1a Queries were formed from the message prompts used in the database recording. To reduce variations in word form, query words were suffix stripped to stems using the standard Porter algorithm [20]. Queries were formed from the prompts by selecting those stems also found in a keyword stem list. 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
```

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 were assumed to be not relevant, even though they are quite likely to contain similar keywords.

VMR1b This is a more realistic set of requests and relevance assessments, collected from the user community that supplied the database messages. A total of 50 requests were collected, 5 for each of the 10 categories used in message collection. These were gathered from 10 users, each of whom 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 categories in that set. For each category a text prompt was formed by combining information given in the 5 message prompts associated with the category.

Users were shown the prompt for the category and asked to compose a natural language request from the information given in the prompt. Users were asked that their request include at least one of the keywords associated with the category.

As for VMR1a request words were suffix-stripped using the Porter algorithm and search queries were formed by selecting the keyword stems. 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

Ideally, the relevance of all archived messages should be assessed; however this is not practical even for our 300 message archive. A suitable assessment subset was formed by combining the 30 messages in the category to which the original message prompt belonged, plus 5 messages from outside the category having the highest query-message scores (computed using collection frequency weighting (see section 6.4.2) on the VMR1 document archive). Subjects were presented with the transcription of each potentially relevant document in random order and asked to mark it as “relevant”, “partially relevant”, or “not relevant”. The average number of highly relevant documents was 10.8, while 17.2 were judged highly or partially relevant. The following sections report results only for the highly relevant relevance set. A full description of the VMR1b naturalistic request set is contained in [21].

Apart from the greater realism, the main difference between VMR1a and VMR1b is that there were far fewer terms per query for the latter, an average of 2.6 distinct terms, against an average of 4.6 for VMR1a.

6.4.2 Query-Document Matching

Document retrieval experiments compared three forms of document scoring: *unweighted* (*uw*) term matching, *collection frequency weighting* (*cfw*), and *combined weight* (*cw*) which takes into account several factors. The unweighted score is simply the sum of matching terms occurring in both the query and the document. The collection frequency is conventional inverse document frequency weighting, computed as

$$cfw(i) = \log \frac{N}{n(i)},$$

where $cfw(i)$ is the cfw weight of term i , N is the total number of documents and $n(i)$ is the number of documents in which term i appears. The combined weight incorporates cfw, within-document term frequency, and normalised document length. The cw weight was defined in [22] and derived in [23]: the cw scheme reflects the City University work for TREC [24]. 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 frequency of i in j , 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 testing we set $K = 1$.

Wt Scheme		Text			Phonetic		
		uw	cfw	cw	uw	cfw	cw
Prec.	5 docs	.26	.30	.30	.25	.29	.31
	10 docs	.22	.25	.27	.22	.24	.27
	15 docs	.19	.21	.24	.19	.21	.23
	20 docs	.17	.19	.21	.17	.18	.20
Av Prec.		.29	.33	.36	.28	.32	.35

Table 3: VMR1a text and phonetic text retrieval performance.

Wt Scheme		Text			Phonetic		
		uw	cfw	cw	uw	cfw	cw
Prec.	5 docs	.34	.35	.34	.34	.35	.35
	10 docs	.28	.31	.29	.29	.32	.31
	15 docs	.26	.30	.30	.26	.30	.30
	20 docs	.24	.28	.28	.25	.28	.28
Av Prec.		.30	.33	.35	.30	.34	.36

Table 4: VMR1b text and phonetic text retrieval performance.

6.4.3 Calibration via text retrieval

Retrieval performance for speech documents can be expected to suffer degradation relative to text documents due to either misses or false alarms. 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.

A particular problem with word spotting is that unrelated acoustic events will often resemble valid keywords. For example, the last part of “hello Kate” is acoustically quite similar to the keyword “locate.” Because even the most accurate acoustic models cannot discriminate between homophones, the output of an ideal word spotter that reports all keyword phone sequences 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.

Table 3 shows retrieval performance for the standard transcribed messages (*text*) and for the ideal phonetic text (*phonetic*) with collection VMR1a and table 4 shows that for VMR1b. It can be seen that introducing cfw weighting gives a substantial improvement in performance over the unweighted case, and cw in turn does better than cfw. For VMR1a the text transcription performs better than the phonetic reference; however the opposite is true for VMR1b. We attribute this phenomenon to stemming inconsistencies between the text transcription and phonetic data. VMR1b in particular is sensitive to this effect due to its very short queries. Due to the small size of the message collection absolute values and observed differences must be treated with caution.

6.4.4 Spoken message retrieval performance

As described previously, the word spotter outputs a list of putative keyword hits and associated acoustic scores. It is found that acoustic false alarms frequently score worse than true hits, and hence a score threshold can be applied to remove most of the false alarms. Clearly, it is desirable 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 [2]). Tables 5, 6 and 7 show spoken document retrieval performance for the three investigated weighting schemes and the different word spotting models at the best *a posteriori* threshold. In practice, an *a priori* fixed threshold would be used in an operational system. For simplicity only average precision values are shown in these tables. It is found that precision at the cutoff values shown in tables 3 and 4 follow similar relative performance trends to the average precision observed for different acoustic model sets.

It can be seen from these tables that talker-dependent models produce the best overall retrieval performance values. This was anticipated from the superior word spotting FOM discussed earlier. Also, supervised

		Av. Prec.	Text (%)	Phon. (%)
Text		.293	100.0	—
Phonetic		.279	95.2	100.0
Head	Depen	.259	88.4	93.8
	Indep	.241	82.4	86.5
	R13	.234	80.1	84.0
	R75	.256	87.6	91.9
Desk	Depen	.241	82.3	86.4
	Indep	.184	62.8	65.9
	R13	.172	58.6	61.5
	R75	.181	61.9	64.9

VMR1a

		Av. Prec.	Text (%)	Phon. (%)
Text		.296	100.0	—
Phonetic		.302	102.0	100.0
Head	Depen	.265	89.5	87.7
	Indep	.249	84.0	82.2
	R13	.245	82.9	81.2
	R75	.270	91.1	89.2
Desk	Depen	.254	85.8	84.1
	Indep	.214	72.3	70.8
	R13	.203	68.6	67.1
	R75	.234	79.2	77.5

VMR1b

Table 5: Retrieval performance with uw scheme.

		Av. Prec.	Text (%)	Phon. (%)
Text		.332	100.0	—
Phonetic		.317	95.5	100.0
Head	Depen	.295	88.8	93.1
	Indep	.263	79.2	83.0
	R13	.274	82.5	86.4
	R75	.294	88.6	92.7
Desk	Depen	.283	85.2	89.3
	Indep	.219	66.0	69.1
	R13	.211	63.4	66.5
	R75	.227	71.6	76.3

VMR1a

		Av. Prec.	Text (%)	Phon. (%)
Text		.332	100.0	—
Phonetic		.339	102.1	100.0
Head	Depen	.312	94.0	92.0
	Indep	.287	86.4	84.7
	R13	.283	85.2	83.5
	R75	.307	92.5	90.6
Desk	Depen	.296	89.2	87.3
	Indep	.249	75.0	73.5
	R13	.251	75.5	73.8
	R75	.276	83.1	81.2

VMR1b

Table 6: Retrieval performance with cfw scheme.

		Av. Prec.	Text (%)	Phon. (%)
Text		.358	100.0	—
Phonetic		.349	97.5	100.0
Head	Depen	.316	88.3	90.5
	Indep	.300	83.8	86.0
	R13	.324	90.5	92.8
	R75	.338	94.4	96.8
Desk	Depen	.299	83.5	85.7
	Indep	.275	76.8	78.8
	R13	.243	67.8	69.5
	R75	.266	74.4	76.3

VMR1a

		Av. Prec.	Text (%)	Phon. (%)
Text		.346	100.0	—
Phonetic		.355	102.6	100.0
Head	Depen	.330	95.4	93.0
	Indep	.309	89.3	87.0
	R13	.303	87.6	85.4
	R75	.335	96.8	94.4
Desk	Depen	.315	91.0	88.7
	Indep	.265	76.6	74.6
	R13	.271	78.5	76.5
	R75	.295	85.3	83.1

VMR1b

Table 7: Retrieval performance with cw scheme.

talker adaptation can significantly improve retrieval performance. In some cases, talker adapted models using R75 adaptation data achieve superior retrieval performance to talker-dependent models. The reason for this is not clear, however, it is important to remember that the document set here is very small. Performance trends for the different weighting schemes are similar to those observed for the text standards, with cw again producing the best performance. Note not only does cw give the best absolute values but performance relative to the text standards is improved as well.

Figure 6: No of putative hits versus threshold for talker-dependent models.

Figure 6 illustrates the effect of increasing score threshold on the number of true hits and false alarms for talker-dependent head and desk microphone models. Figure 7 shows spoken document retrieval performance for VMR1a and VMR1b using talker-dependent head-microphone models at different acoustic thresholds for uw, cfw and cw schemes. These figures show that the performance trends observed for the *a posteriori* best performance thresholds are consistent across the different threshold levels. Also, significantly, 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. This trend is more pronounced with the VMR1a results which is probably due to their longer average query length.

Figure 8 shows that increasing the keyword transition penalty (shown in figure 5) dramatically reduces the number of false alarms, while only mildly impacting the number of correctly identified keywords. Figure 9 shows the effect of keyword score thresholding for word spotting output from systems with different keyword transition penalties. It can be seen that there is little difference between the best retrieval performance at the optimal *a posteriori* threshold. However, the system with fewest false alarms prior to the use of thresholding (transition penalty 100) not only exhibits the best available individual retrieval performance, but is also less sensitive to variation in the acoustic score threshold.

6.4.5 Retrieval results summary

In summary, the following general points can be made:

- Term weighting schemes developed for text retrieval transfer well to the retrieval of spoken document.
- Spoken document retrieval performance of between 75% and 95% of that achieved with text transcriptions can be obtained depending on the generality of the acoustic recognition models.
- Unsurprisingly, spoken document retrieval performance is adversely affected by degradation in word spotting performance.

VMR1a

VMR1b

Figure 7: Average retrieval performance versus threshold for talker-dependent head-microphone models.

Figure 8: False alarms vs. keyword transition penalty (*R75* models).

Figure 9: Relative average precision vs. keyword score threshold for VMRIa using (*R75* models) with keyword transition penalty varied between 20 and 100.

6.5 User Interface

We have developed a prototype VMR application that integrates keyword spotting, information retrieval, and video capture/playback capabilities. The audio soundtrack of each message (whether from an existing archive or received as new mail) is passed to the acoustic word spotter. This computes a sequence of putative keyword hits, which is added to an index containing all putative hits for all messages, along with mail header information and pointers to the message data (for playback). Because the computationally-intensive word spotting phase is done off-line (as messages are added to the archive), retrieval of archived messages is nearly instantaneous.

Figure 10: Video Mail User Interface application.

The VMR user interface is shown in figure 10. The interface shows a scrollable list of all available messages in the user's video mail archive. Various controls let the user "narrow" the list, for example, by displaying only those messages from a particular user or received after a particular time. Unsetting a constraint restores the messages hidden by that constraint; multiple constraints can be active at one time, giving the messages selected by a boolean conjunction of the constraints. With no constraints messages are ranked by origination date, such that the most recently received document is displayed at the top of the list. When the user inputs a search query the retrieval engines computes the resulting score for each message. The interface then displays a list of messages ranked by score, with the scores shown as bar graphs. Messages with identical scores are ranked by date.

To review an individual message, a "video browser" can be activated. The video browser, shown in figure 11, graphically represents the message as a dark horizontal bar, with putative keyword hits displayed as lighter regions. Time runs from left to right, and keyword hits are displayed proportionally to when they occur in a message; for example, keywords at the beginning appear on the left side of the bar. The brightness of a keyword region is proportional to the score computed by the word spotter, so that more likely hits appear brighter and stand out. Portions of the video message can be selected for playback by dragging

Figure 11: Prototype mail browser.

over part of the bar; this enables the user to selectively audition regions of interest rather than the entire message. A more detailed discussion of the system is contained in [1].

7 Conclusions and Future Work

The results obtained in the VMR project so far suggest that spoken document retrieval is a feasible proposition. Retrieval performance of between 75% and 90% of text has been achieved for indexing via word spotting. Of course, the ultimate evaluation is whether users find the retrieval tool useful in day-to-day operation.

Next stage system development in VMR will concentrate on greater integration with the ORL Medusa system and enhancing the user interface, hopefully including including input from real users.

The speech recognition component is currently being extended to use of a large vocabulary recognition system, and also a phone lattice scan approach developed by James [25] to search for out-of-vocabulary query terms.

Further directions for the project include the development of a system for the automated retrieval of broadcast TV news. Initial results from this work using text subtitles to index the data are very encouraging [26] and we intend to extend this to indexing using speech recognition in the near future.

8 Acknowledgements

This project is supported by the UK DTI Grant IED4/1/5804 and EPSRC Grant GR/H87629. Olivetti Research Limited is an industrial partner of the VMR project and we are indebted to them for the use of the Medusa system. We wish to thank Julian Odell for the baseline triphone models, and Chris Leggetter and Phil Woodland for the talker-adaptation software.

Details of the VMR project are contained in the project web page:

<http://svr-www.eng.cam.ac.uk/Research/Projects/VideoMailRetrievalVoice/>

from which copies of VMR publications can be obtained.

References

- [1] M. G. Brown, J. T. Foote, G. J. F. Jones, K. Spärck Jones, and S. J. Young. Video Mail Retrieval using Voice: An overview of the Cambridge/Olivetti retrieval system. In *Proc. ACM Multimedia 94 Workshop on Multimedia Database Management Systems*, pages 47–55, San Francisco, CA, October 1994.
- [2] G. J. F. Jones, J. T. Foote, K. Spärck Jones, and S. J. Young. Video Mail Retrieval: the effect of word spotting accuracy on precision. In *Proc. ICASSP 95*, pages 309–312, Detroit, May 1995. IEEE.
- [3] J. T. Foote, G. J. F. Jones, K. Spärck Jones, and S. J. Young. Talker-independent keyword spotting for information retrieval. In *Proc. Eurospeech 95*, volume 3, pages 2145–2148, Madrid, 1995. ESCA.
- [4] K. Spärck Jones, G. J. F. Jones, J. T. Foote, and S. J. Young. Retrieving spoken documents: VMR project experiments. Technical Report 366, Cambridge University Computer Laboratory, May 1995.
- [5] U. Glavitsch, P. Schäubel, and M. Wechsler. Metadata for integrating speech documents in a text retrieval system. *SIGMOD Record*, 23(4):57–63, December 1994.
- [6] M. Wechsler and P. Schäubel. Indexing methods for a speech retrieval system. In Ian Ruthven, editor, *Proceedings of the MIRO Workshop*, University of Glasgow, September 1995.
- [7] S. Wray, T. Glauert, and A. Hopper. The Medusa applications environment. In *Proceedings IEEE International Conference on Multimedia Computing and Systems*, pages 265–273, Boston, May 1994. IEEE.
- [8] A. Hopper. Digital video on computer workstations. In *Proceedings of Eurographics*, 1992.
- [9] I. Leslie, D. McAuley, and D. Tennenhouse. ATM Everywhere? *IEEE Network*, March 1993.
- [10] J. K. Ousterhout. *Tcl and the Tk Toolkit*. Addison-Wesley, 1994.
- [11] M. Sanderson. Word sense disambiguation and information retrieval. In *Proc. SIGIR 94*, Dublin, 1994. ACM, Springer-Verlag.
- [12] G. J. F. Jones, J. T. Foote, K. Spärck Jones, and S. J. Young. VMR report on keyword definition and data collection. Technical Report 335, Cambridge University Computer Laboratory, May 1994.
- [13] L. F. Lamel, H. K. Kassel, and S. Seneff. Speech database development: Design and analysis of the acoustic-phonetic corpus. In *Proc. DARPA Speech Recognition Workshop*, pages 26–32, 1986.
- [14] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals. WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition. In *Proc. ICASSP 95*, pages 81–84, Detroit, May 1995. IEEE.
- [15] R. C. Rose and D. B. Paul. A hidden Markov model based keyword recognition system. In *Proc. ICASSP 90*, pages 129–132. IEEE, 1990.
- [16] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286, February 1989.
- [17] S. J. Young, P. C. Woodland, and W. J. Byrne. *HTK: Hidden Markov Model Toolkit V1.5*. Entropic Research Laboratories, Inc., 600 Pennsylvania Ave. SE, Suite 202, Washington, DC 20003 USA, 1993.
- [18] S. J. Young, J. J. Odell, and P. C. Woodland. Tree-based state tying for high accuracy acoustic modelling. In *Proc. ARPA Spoken Language Technology Workshop*, Plainsboro, NJ, 1994.
- [19] C. J. Leggetter and P. C. Woodland. Flexible speaker adaptation for large vocabulary speech recognition. In *Proc. Eurospeech 95*. ESCA, 1995.
- [20] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, July 1980.

- [21] G. J. F. Jones, J. T. Foote, and K. Spärck Jones. Video Mail Retrieval using voice: Report on collection of naturalistic requests and relevance assessments. VMR Project Working Document, April 1995.
- [22] S. E. Robertson and K. Spärck Jones. Simple, proven approaches to text retrieval. Technical report, Cambridge University Computer Laboratory, Dec 1994.
- [23] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proc. SIGIR 94*, pages 232–241, Dublin, 1994. ACM.
- [24] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In D. K. Harman, editor, *The Third Text REtrieval Conference (TREC-3)*, pages 109–126. 1995.
- [25] D. A. James and S. J. Young. A fast lattice-based approach to vocabulary independent wordspotting. In *Proc. ICASSP 94*, volume I, pages 377–380, Adelaide, 1994. IEEE.
- [26] M. G. Brown, J. T. Foote, G. J. F. Jones, K. Sparck Jones, and S. J. Young. Automatic content-based retrieval of broadcast news. In *Proc. ACM Multimedia 95*, pages 35–43, San Francisco, November 1995. ACM.