Technical Report

UCAM-CL-TR-402
ISSN 1476-2986

Number 402

UNIVERSITY OF
CAMBRIDGE

Computer Laboratory

Video mail retrieval using voice: report on collection of naturalistic requests and relevance assessments

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

September 1996

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

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

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

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

ISSN 1476-2986

Video Mail Retrieval Using Voice : Report on Collection of Naturalistic Requests and Relevance Assessments *

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

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

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

September 1996

Abstract

This report describes the rationale, design, collection and initial statistics of a message request and retrieved document relevance assessment set for the Cambridge Video Mail Retrieval (VMR) Project. This data set is designed to complement the VMR Database 1 (VMR1) message set described in [?] and was designed for the testing of document searching methods being investigated in the VMR project. The combined message and request set is referred to as VMR1b.

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

1 Introduction

This report describes the motivation, design and collection of a set of naturalistic message requests and corresponding relevance assessments for the Cambridge University (Engineering Department (CUED) and Computer Laboratory (CUCL)), and Olivetti Research Limited (ORL) research project on Video Mail Retrieval (VMR) [Hopper et al., 1993]. This set of requests and assessments is complementary to the VMR Database 1 spoken message archive collected in Stage 1 of the VMR project. A full description of Database 1 (VMR1) is contained in [Jones et al., 1994] and is beyond the scope of this report.

The following brief outline summaries the principal features of VMR1 relevant to the current report. VMR1 was designed to meet various criteria arising from the requirements of both the retrieval aspects and the speech recognition (word spotting) requirements of the project. Recorded messages needed to have similar general properties to those anticipated in an operational video mail installation. An additional requirement was that the messages make natural use of the set of fixed keywords used for the Stage 1 word spotting investigations. The message achieve is very small by IR standards and hence had to be carefully structured. Messages were sought on *topics* within a set of topic *categories*: each category has an associated set of *keywords* drawn from the fixed keyword *vocabulary* from which search *terms* in Stage 1 must be taken. Since they did not come 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. Apart for the message set, VMR1 also contains spoken material for use as training data for speaker-dependent acoustic models which were used to recognise the occurrence of search terms in the speech messages during Stage 1 of the VMR project.

The total keyword vocabulary was 35 words, along with a set of 31 related *otherwords*. These were divided between a set of 10 broad subject categories, where appropriate keywords and otherwords were assigned to more than one category. The list of categories is shown in Appendix A. There were 5 scenario prompts provided for each category. 20 messages from a subset of 4 categories were recorded by each of 15 speakers giving 300 messages in total. For any one category there were messages from 6 speakers. The assignment of speakers to categories was randomised, 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 in the form 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 acoustic keyword 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 category subsets 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]. The same technique was used in the distribution of categories in the collection of message requests and is described in section 2.4.

Early IR experiments in the VMR project concentrated on a basic comparison of text and spoken document retrieval performance [Jones et al., 1995]. The queries were generated from the existing scenarios, and messages were assumed relevant to (and only to) the prompt which suggested them. This type of query and relevance judgement set does not stand up to scrutiny and hence the purpose of this request collection was to generate a set of more realistic requests and relevance assessments.

2 Specification of Data Collection Task

Message requests and corresponding relevance assessments are an important part of information retrieval research exercises. Within the scope of the VMR project these are defined as follows.

Requests

A request is a concise expression of a user's information need; assuming that the user suitably expresses their actual need in the request statement. A request can be processed in various ways to form a query. This query can be used to seek potentially relevant documents in the available message archive.

Typically a request might take the form of a free vocabulary natural language sentence. However in Stages 1 and 2, the VMR project was limited to its fixed retrieval vocabulary and hence, our requests had to contain sufficient instances of these fixed search terms to be useful. Our requests are thus open vocabulary natural language sentences with a bias towards our fixed keyword vocabulary. The collection of requests in this form meant that they are useful for later experiments when open search term vocabularies are allowed and, incidentally, that the information search of the originator can be more nearly (if not completely) understood directly from the request.

Relevance Assessments

For each request the originator should in theory assess the relevance of all messages in the archive to each request they generate. In practice, however, even for the VMR corpus containing only 300 documents, this is impractical. A generally accepted approximation is to pool the documents retrieved at a certain rank level by a number of different strategies and assess the relevance of all members of the pooled set. The assumption here is, of course, that all (or hopefully nearly all) potentially relevant documents will be present in this set.

Another factor which needs to be taken into account for relevance assessments is sequencing effects between different messages if they are presented to more than one assessor in the same order. This difficulty arises because, despite attempting to do so, it is not possible for an assessor to treat each message entirely independently of the others seen for a particular request.

2.1 Request Archive Organisation

The first VMR request set consists of the following:

- Total 50 natural language requests.
- 5 requests for each of the 10 categories.
- 5 requests each originated by 10 individuals, each one generating requests for a unique sequence of 5 categories.

2.2 Request Motivation

For each category the 5 message collection scenario prompts were manually combined to form a short category summary which could be used as a message request motivation paragraph.

Example

As an example of this procedure consider the combination of the specifications of the debugging category *word processing* into an request motivation paragraph.

The keywords, otherwords and scenarios for this category are as follows:

```
CAT-DEBUG WORD PROCESSING (WP)

   KEYWORDS (WP-KW)
      latex  spellcheck  document  edit

   OTHERWORDS (WP-OW)
      postscript  emacs  format  paper

   SCENARIOS

      WP-SC-1 Send a message describing how you might go about generating
              figures in latex. Which of these methods would allow me to
              shade the diagrams for clarity ?

      WP-SC-2 I'd like to get some idea of the performance one can
              expect from spellcheck programs. If you've got a
              document with lots of figures and tables what effect
              might this have on the performance figures for the
              spellchecker ?

      WP-SC-3 How convenient is it to send documents round as postscript
              files via email, for example are they too bulky for the
              network ? I've several project reports which about 10 people
              are interested in seeing, should I contemplate sending them the
              postscript direct.

      WP-SC-4 I'm sending you the draft version of the project report. Please
              mail me with your comments about style, content and accuracy.
              Let me know if you want to edit it yourself and when you might
```

be able to do it.

WP-SC-5 I've got to do a workshop submission with some rather unusual features, do you know where I might be able to find some non standard style files which I might be able to use ?

Using this information the following motivational paragraph was composed.

REQUEST MOTIVATING PARAGRAPH

Word processing of documents in Latex requires knowledge of several distinct subject areas. For example, the formatting of documents to defined standards, checking the accuracy of spelling and punctuation, and composition and inclusion of relevant figures and tables. Documents can be written by authors at multiple sites and in this case users must be able to exchange both text and formatted files in the development and checking of their work. These messages concern questions about issues relating to these skills, for example, how to automatically check the grammar of text document.

The complete set of scenarios for all categories is given in [Jones et al., 1994] and the request motivation paragraphs are shown in Appendix B.

2.3 Request Collection

The request originator was shown a motivational paragraph together with the fixed keywords and otherwords assigned to the category. In response to this they were asked to write a natural language request of a sentence or two, using search terms which they hoped would retrieve messages on some aspect of the expected category content material, as suggested by the motivating paragraph. They were asked that this request should contain at least one of the fixed keywords associated with the category. The request might actually contain the keywords or otherwords associated with one or more other categories, but since this is a potential operational retrieval impairment these additional keywords were not excluded from the requests. Users were told they could use the keywords in whatever form they wish. For example, the keyword *manage* is equally acceptable in the forms, *management*, *manager* or *managing*.

It is likely that request originators will not have chosen subject matter from the category paragraphs in any sort of even distribution. This means that some messages will never be judged relevant to a request in this set. However although we could have sought to resolve this situation we decided not to since it is probably representative of an operational system that may contain mail which is not required in current searches.

2.4 Category sequencing

In order to remove sequence effects between categories it was necessary to impose suitable distributions.

Latin Square Sequences A suitable scheme for this distribution can be derived using the procedure of Latin squares. A Latin square is an n by n table or array in which the entries are n distinct symbols, assigned so that each appears once in each row and in each column. For example, a 3 by 3 Latin square could have either of the following forms:

1	2	3
2	3	1
3	1	2

1	3	2
3	2	1
2	1	3

(from [Tague, 1981], page 79).

Applying this idea to the sequencing of the motivation paragraphs the following category sequences are formed.

Let categories given to each assessor be A B C D E.

	A	B	C	D	E
1	C1	C2	C3	C4	C5
2	C2	C1	C4	C6	C7
3	C3	C10	C5	C7	C6
4	C4	C9	C6	C8	C10
5	C5	C8	C7	C9	C4
6	C6	C7	C8	C10	C3
7	C7	C6	C9	C1	C2
8	C8	C5	C10	C2	C1
9	C9	C4	C1	C3	C8
10	C10	C3	C2	C5	C9

Since we have less categories per assessor than distinct categories, only 5 categories appear in each row.

Appendix C shows the ordered category groups formed using this technique for the request collection.

For the original collection of messages categories were, as far as possible, only supplied to speakers who were familiar with the subject of the category. It was felt sensible to adopt a similar policy here, however it is probably not quite so important here since request originator's were generally given sufficient motivational material along with the category's keywords enabling them to form meaningful requests on subjects about which they were not particularly knowledgeable.

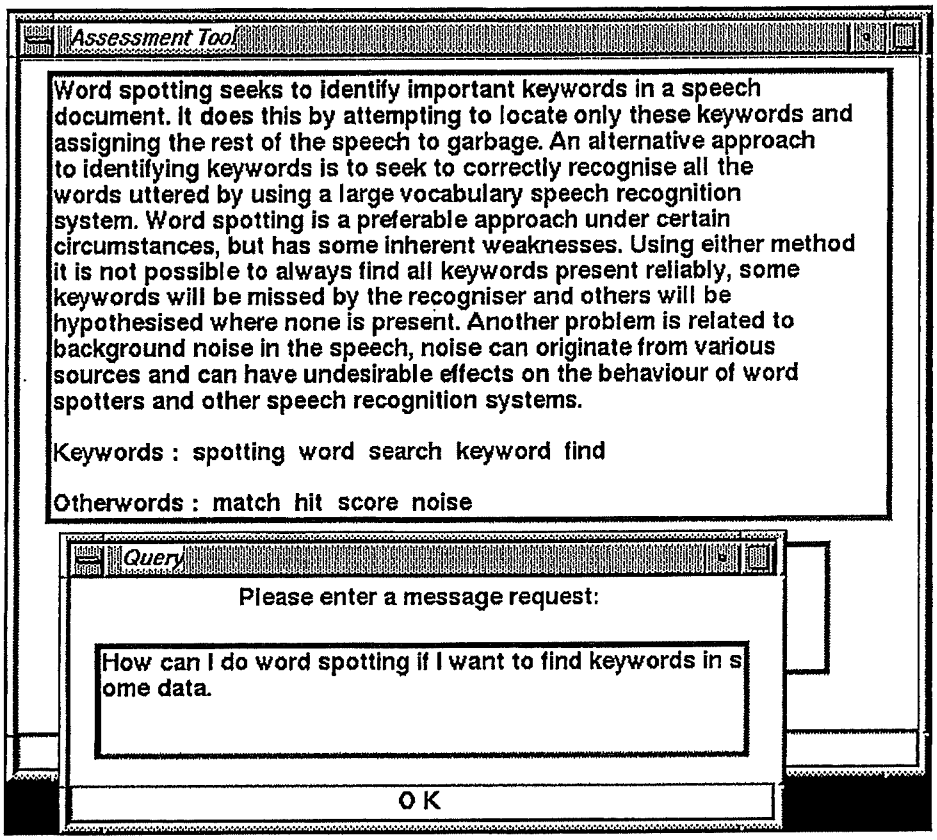
2.5 Relevance Assessment Collection

Since they were only being asked to generate 5 requests, it was felt reasonable to ask the request originators' to assess the relevance of retrieved messages for each of their requests.

Also, since we had decided not to require them to assess the relevance of all messages in the archive to each request, we needed to design a suitable message subset or pool to show them.

It was decided, as a practical compromise, that for each request the user should assess:

1. the 30 messages that were recorded for the category associated with the request.



Word spotting seeks to identify important keywords in a speech document. It does this by attempting to locate only these keywords and assigning the rest of the speech to garbage. An alternative approach to identifying keywords is to seek to correctly recognise all the words uttered by using a large vocabulary speech recognition system. Word spotting is a preferable approach under certain circumstances, but has some inherent weaknesses. Using either method it is not possible to always find all keywords present reliably, some keywords will be missed by the recogniser and others will be hypothesised where none is present. Another problem is related to background noise in the speech, noise can originate from various sources and can have undesirable effects on the behaviour of word spotters and other speech recognition systems.

Keywords : spotting word search keyword find

Otherwords : match hit score noise

Please enter a message request:

How can I do word spotting if I want to find keywords in some data.

Figure 1: Request collection user interface.

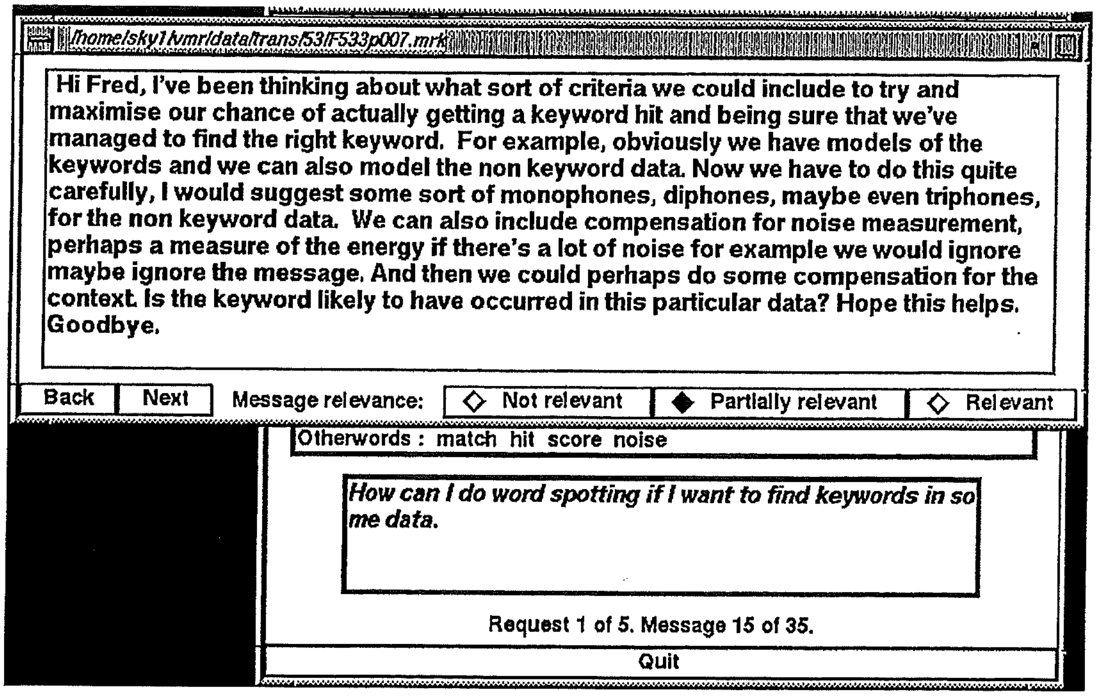
2. the 5 highest scoring messages *not* in this category. These messages will hopefully cover any relevant to the request which belong to different categories. These 5 messages were chosen from the ranked message list formed by scoring all messages in the archive using a *query* generated from the originator's request. The query contains the suffix stripped [Porter, 1980] elements of the request which match suffix stripped elements from the fixed keyword list. Similarly the document vectors used were composed of suffix stripped fixed keywords which occurred in the full manual transcription of the messages. A standard query–document matching score using collection frequency weighting (cfw) (or inverse document frequency weighting) scheme [Robertson and Sparck Jones, 1994].

Thus, for each request, each user assessed the relevance of 35 messages.

If the 30 messages for each category were always presented to the user in the same order, it is possible that sequencing effects may occur in the assessment of relevance. Therefore for each request the 35 messages were sorted into a pseudo random ordered list for assessment.

2.6 Request and Assessment Tool

A graphical request and relevance assessment collection tool was written using the Tcl/Tk scripting language [Ousterhout, 1994]. Users were shown instructions and an appropriate



Hi Fred, I've been thinking about what sort of criteria we could include to try and maximise our chance of actually getting a keyword hit and being sure that we've managed to find the right keyword. For example, obviously we have models of the keywords and we can also model the non keyword data. Now we have to do this quite carefully, I would suggest some sort of monophones, diphones, maybe even triphones, for the non keyword data. We can also include compensation for noise measurement, perhaps a measure of the energy if there's a lot of noise for example we would ignore maybe ignore the message. And then we could perhaps do some compensation for the context. Is the keyword likely to have occurred in this particular data? Hope this helps. Goodbye.

Otherwords : match hit score noise

How can I do word spotting if I want to find keywords in some data.

Request 1 of 5. Message 15 of 35.

Figure 2: Message relevance assessment user interface.

motivating paragraph and then asked to enter their request. Figure 1 shows the request collection interface.

The request was processed and the user was then shown the text transcription of each of the sequence of 35 messages to assess for relevance. Each message was marked as being "relevant", "partially relevant" or "not relevant" to the request. Figure 2 shows the relevance assessment interface. After marking the current message the system moves on to the next message, however the user was allowed to return to previous messages to check or change previously marked relevance assessments.

2.7 Request and Assessment Collection Procedure

The collection procedure was divided into sessions. One session constituted the complete collection for each user. Hence, there were 10 sessions in all, each consisting of five prompts.

2.7.1 Session format

The welcome and instruction prompts used are shown in Appendix D.

```
- welcome prompt

- for ( x = 0; x < no_of_prompts; x++ )

  {

    - motivating paragraph (x)

    - collect request
  - convert to query (suffix strip items, compare to elements of
                      suffix stripped fixed keyword list)

  - compute vector product score for the query and each cfw weighted
    transcribed message

  - rank messages by score and find the five highest scoring messages which
    are not in the category of the prompt

  - compose complete list of 35 potentially relevant items and randomise
    order of list

  - for ( y = 0; y < 35; y++ )

      - show message transcription(y) to user and get relevance assessment

  - form list of relevant documents for this request

}
```

Once all the requests and assessments had been gathered, they were combined into a single file TREC relevance file [Salton, 1991]. In order to be able to manage the requests each was assigned a unique ID. The derivation of request ID's is explained in Appendix E.

3 Overview of Requests and Relevance Assessments

3.1 Basic Statistics

	Min	Av	Max
No of words in each request	5	12.00	25
No of standard search terms in each request	3	7.84	15
No of suffix stripped fixed keywords in each request	1*	2.68	5

* – one request where the user had not obeyed the instructions carefully was found to contain no fixed keywords.

Table 1: Summary of request statistics.

Table 1 summarises the statistics of the natural language requests. This also shows the details measured in terms of members of the fixed keyword list included in the requests and of standard search terms used. As described previously, fixed keywords are identified by comparing suffix stripped elements from the request with the suffix stripped members of the fixed keyword list. The set of standard search terms for each request is formed by removing members of the standard van Rijsbergen *stop list* from the request [van Rijsbergen, 1979]. Processing of the requests in this way corresponds to formation of normal open search term queries.

Table 2 shows a short summary of the message relevance assessments for the requests.

	Min	Av	Max
No of "relevant" messages per request	0	10.80	31
No of "partially relevant" messages per request	1	17.22	35

Table 2: Summary of relevance assessment statistics.

Weight Scheme		35 Keywords		
		uw	cfw	cw
Precision	5 docs	0.342	0.350	0.342
	10 docs	0.281	0.308	0.294
	15 docs	0.260	0.297	0.299
	20 docs	0.242	0.281	0.280
Av Precision		0.296	0.332	0.346

(a) using only 35 fixed keywords.

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

(b) using open search vocabulary.

Table 3: Experimental retrieval results using naturalistic request/relevance set and message transcriptions.

3.2 Retrieval Experiments

Table 3 shows experimental retrieval results using the naturalistic request/relevance set with the full message transcriptions. Results are shown for retrieval experiments with the 35 fixed keyword vocabulary and open search vocabulary using all content words. Only messages marked highly relevant by users are considered relevant in this experiment. Term weighting applied in each case is as follows: *uw* unweighted (coordination), *cfw* collection frequency weighted (inverse document frequency), and *cw* combined weighted as defined in [Robertson and Sparck Jones, 1994] with K=1. The results shown here show anticipated performance trends. Retrieval performance is improved by using progressively more complex term weighting schemes. Use of all search terms gives superior performance to that of a small fixed keyword vocabulary. This difference would probably be more appreciable were the fixed keywords used not so well matched to the contents of the message archive.

4 Concluding Remarks

We believe that the requests and relevance assessment set gathered in this data collection exercise are sufficient to perform useful retrieval experiments for our spoken message archive. However, whilst observed experimental behaviour is interesting all results must be treated carefully and trends considered indicative rather than proven due to the small size of the message set.

References

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

[Jones et al., 1994] Jones, G. J. F., Foote, J. T., Sparck Jones, K., and Young, S. J. (1994). Video Mail Retrieval Using Voice: Report on keyword definition and data collection. Technical Report 335, Cambridge University Computer Laboratory.

[Jones et al., 1995] Jones, G. J. F., Foote, J. T., Sparck Jones, K., and Young, S. J. (1995). Video Mail Retrieval: the effect of word spotting accuracy on precision. In *Proceedings of ICASSP 95*, Detroit. IEEE.

[Ousterhout, 1994] Ousterhout, J. K. (1994). *Tcl and the Tk Toolkit.* Addison-Wesley.

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

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

[Salton, 1991] Salton, G. (1991). TREC. Software distribution.

[Tague, 1981] Tague, J. M. (1981). The pragmatics of information retrieval experimentation. In Sparck Jones, K., editor, *Information Retrieval Experiment*, chapter 5, pages 59–102. Butterworths.

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

A Categories Definition

The VMR database 1 categories are defined as follows:

1: Spotting

2: Document

3: Output

4: Retrieval

5: Windows

6: Management

7: Badge

8: Pandora

9: Schedule

10: Equipment

B Category Message Request Prompts

The following are the request prompts formed for each category by combining the original message scenario prompts given in [Jones et al., 1994].

CAT-1 SPOTTING (SP)

SP-RP

Word spotting seeks to identify important keywords in a speech document. It does this by attempting to locate only these keywords and assigning the rest of the speech to garbage. An alternative approach to identifying keywords is to seek to correctly recognise all the words uttered by using a large vocabulary speech recognition system. Word spotting is a preferable approach under certain circumstances, but has some inherent weaknesses. Using either method it is not possible to always find all keywords present reliably, some keywords will be missed by the recogniser and others will be hypothesised where none is present. Another problem is related to background noise in the speech, noise can originate from various sources and can have undesirable effects on the behaviour of word spotters and other speech recognition systems.

Keywords : spotting word search keyword find

Otherwords : match hit score noise

CAT-2 DOCUMENT (DO)

DO-RP

It is important for receiver's of speech and video mail messages to have a clear plan for the storage and deletion of their messages. This is particularly important if much received mail is either irrelevant to a user or unsolicited junk. Employers can request staff to manage their file space efficiently, to limit the use of disk space; however this is difficult to enforce.

There are advantages to sending mail documents rather than more informal communication, and they can be made easier to manage by the use of short content indicating tags.

Keywords : document message mail

Otherwords : item file video_mail

CAT-3 OUTPUT (OP)

OP-RP

The presentation format of the output from any message retrieval system is important. This is particularly true in the case of multi media systems. In multimedia environments there is much visual information available which may be of use to operators in assessing the relevance of a message to their enquiry.

The order in which documents are presented can assist users in deciding which messages to examine. Hence, the factors used to derive the score which determines this ranking are important.

Keywords : output score rank assess

Otherwords : list

CAT-4 RETRIEVAL (RE)

RE-RP

The search procedure in a document retrieval involves several stages: specification of a request for messages using keywords, dates, users, etc; browsing retrieved messages looking for the search target(s); and possibly using the retrieved material to help specify a modified request using relevance feedback. The details used in requests can affect the behaviour of the retrieval system. Hence it is desirable to form efficient queries to help you find the documents you are looking for, with the minimum of effort.

Keywords : retrieve search find display

Otherwords : retrieval browse listen get

CAT-5 WINDOWS (WI)

WI-RP

Windows based interfaces are available on all current workstations. These have many uses and can potentially make the working environment very efficient. Windows is especially appropriate for multimedia systems which can generate graphical output from multiple sources. The windows environment is very flexible enabling each user to design a personalised interface with menus, icons and windows of their choice. Of course, not all software is easy to configure for different machines and users can encounter operational problems if things are not set up correctly.

Keywords : windows display interface

Otherwords : X-windows screen graphics server

CAT-6 MANAGEMENT (MA)

MA-RP

Communications concerning management related matters can encompass many subjects. For example, messages from staff to their boss regarding the progress of a project, or perhaps communication of research results and their implications for the project plan. Other subjects might concern the formation of project plans, alterations to plans resulting from failure to deliver required components, and the scheduling of management meetings.

Keywords : manage project meeting plan

Otherwords : management work schedule task

CAT-7 BADGE (BA)

BA-RP

The active badge system is very useful for quickly locating staff and provides an efficient method of office security (as long as you remember to take your active badge out of the building with you). Inevitably this operational environment is somewhat compromised if sensors fail unexpectedly. The potential of systems like the active badge are still largely unrealised and many additional applications could become available if they were integrated into a distributed multimedia environment; although an attempt to do this would inevitably be slowed by solution of technical obstacles.

Keywords : badge active sensor locate location

Otherwords : system xab active-badge

CAT-8 PANDORA (PA)

PA-RP

The introduction of interactive multimedia workstations such as Pandora into offices gives staff access to many new facilities. For instance video mail is a user friendly method of communication; however there can be problems with locating messages if they are not well labelled. There are some points of concern arising from the adoption of a multimedia working environment. For example, some staff feel the presence of cameras permanently directed at them results in a lack of privacy. In practice, these systems are still a relative novelty and we find that visitors are always keen to see them in operation.

Keywords : Pandora workstation camera video mail

Otherwords : speak microphone

CAT-9 SCHEDULE (SC)

SC-RP

The scheduling of tasks to meet certain deadlines is the cause of much interaction between staff. Enquiries are frequently made about the date and time by which something must be completed or delivered, or when the next meeting about a certain project is scheduled to take place. Other communications concern the broadcasting of new or revised schedule information.

Keywords : time date meeting staff

Otherwords : task deadline effort

CAT-10 EQUIPMENT (EQ)

EQ-RP

Workstation networks provide a powerful computing environment. However, users frequently encounter small problems which they require expert advice to sort out. For example indigo workstations allow direct connection of microphones for audio input, but if these do not work as required, or correctly in combination with other input devices, such as the keyword, the user will often need to ask for help.

The ability to use a particular workstation while someone else runs a background process on the same machine enables resources to be used efficiently. However, these background processes can slow down a machine and you may need to restrict other users' access to your machine in order to meet research deadlines.

Keywords : indigo workstation microphone network

Otherwords : keyboard unix display disc

C Assignment of Groups of Categories to Each Operator

Taking into account the comments about assignment of categories to operators who would hopefully be able to form requests and make relevance assessments with some knowledge, the following category assignments were made. These were chosen on the basis of trying to place together categories Active Badge and Pandora as ORL type categories and Spotting and Output as CUED type categories. The more general ones were then used to fill the gaps in the operator groups. Despite this apparent restriction the category distribution is still observed to be quite broad.

C1 -	spotting
C2 -	document
C3 -	output
C4 -	retrieval
C5 -	windows
C6 -	management
C7 -	badge
C8 -	Pandora
C9 -	schedule
C10 -	equipment

For the individual operators the following subject groups were produced:

OPR1 :	C1	spotting
	C3	output
	C2	document
	C6	management
	C5	windows
OPR2 :	C3	output
	C1	spotting
	C6	management
	C4	retrieval
	C10	equipment
OPR3 :	C2	document
	C9	schedule
	C5	windows
	C10	equipment
	C4	retrieval
OPR4 :	C6	management
	C8	Pandora
	C4	retrieval
	C7	badge
	C9	schedule
OPR5 :	C5	windows
	C7	badge
	C10	equipment
	C8	Pandora
	C6	management

```
OPR6   : C4  retrieval
         C10  equipment
         C7  badge
         C9 schedule
         C2  document

OPR7   : C10  equipment
         C4  retrieval
         C8  Pandora
         C1  spotting
         C3  output

OPR8   : C7  badge
         C5  windows
         C9 schedule
         C3  output
         C1  spotting

OPR9   : C8  Pandora
         C6  management
         C1  spotting
         C2  document
         C7  badge

OPR10  : C9 schedule
         C2  document
         C3  output
         C5  windows
         C8  Pandora
```

Each operator is assigned a number appropriate to their knowledge of the categories.

D Information used to guide operators

Before the exercise began operators were given the following general background information.

General Introduction

Requesting and Relevance Assessment of Transcribed Spoken Documents

Thank you for agreeing to take part in this relevance assessment exercise. In it you will be asked to decide whether each of a series of messages shown to you, is RELEVANT to your (hypothetical) need for messages about a certain topic. The messages are taken from a transcribed spoken message database.

In order to retrieve a potentially relevant set of messages, you will first be asked to express your need for information; these should be natural language sentences.

You will be asked to form a request for each of five message categories. For each category you will first be shown a paragraph summarising the subjects covered within the present category.

You will then be asked to enter a request for messages on a topic covered in this paragraph.

After you have done this a separate program will find the messages which match your request most closely. You will then be shown each of these messages and asked to assess its relevance to your request.

Motivate Request

To motivate a suitable natural language request for each category, operators were given the following prompt.

Requesting Messages

The following paragraph summarises the subjects covered by a number of spoken mail messages. Please compose a natural language sentence or two to act as a request for retrieval of messages about one of the topics. Your request could range from being quite general in nature through to addressing your need for information about a very specific issue.

Associated with the messages are a number of keywords, please make sure your request uses one or more of these to express your requirements. A number of otherwords are listed, one or more of which you may find useful in composing your request. The keywords and otherwords can be used in your request in any form, eg document, documents, documentation, documented.

The following is an example prompt,

> Word processing of documents in Latex requires knowledge of several

distinct subject areas. For example, the formatting of documents to defined standards, checking the accuracy of spelling and punctuation, and composition and inclusion of relevant figures and tables. Documents can be written by authors at multiple sites and in this case users must be able to exchange both text and formatted files in the development and checking of their work. These messages concern questions about issues relating to these skills, for example, how to automatically check the grammar of text document.

Keywords : latex spellcheck document edit
Otherwords : postscript emacs format paper

In response to which you might enter a request of the form,

I'd like to know about document processing with latex.

or,

Give me messages on inclusion of postscript figures in latex documents.

Motivate Assessment

Before assessing the relevance of a series of messages to their request operators were given the following details.

Assessing Message Relevance

A number of messages retrieved in response to your request will now be displayed. Please consider the relevance of each message to your request and select the appropriate button to mark it as "relevant', "partially relevant' or "not relevant". Messages should be marked independently for relevance even if they repeat information from previous messages. If you change your mind occasionally about the relevance of a particular message you can go back to it using the "back" button and return to the current message using the "next" button.

E Request and Relevance Assessment File Format Specifications

The unique request file_id also serves as the filename root for the relevance data and other information files. The filename root is an 8-character string composed of the following fields:

```
rccnnsmm   where

r      : it is related to a request

cc     : 2-digit category no, as shown in Appendix A

nn     : 2-digit unique user id

s      : 1-digit unique collection session id (0 in this collection)

mm     : 2-digit unique request id for this session
```

For example, r0885002 is a request based on category 8 from user 85 in session 0 and is the second request originated in this session.