Technical Report

UCAM-CL-TR-335
ISSN 1476-2986

Number 335

UNIVERSITY OF CAMBRIDGE

Computer Laboratory

Video mail retrieval using voice: report on keyword definition and data collection (deliverable report on VMR task No. 1)

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

April 1994

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

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

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

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

ISSN 1476-2986

Video Mail Retrieval Using Voice * : Report on Keyword Definition and Data Collection (Deliverable Report on VMR Task No 1)

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

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

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

April 1994

Abstract

The report describes the rationale, design, collection and basic statistics of the initial training and test database for the Cambridge Video Mail Retrieval (VMR) Project. This database is intended to support both training for the wordspotting processes and testing for the document searching methods using these that are being developed for the project's message retrieval task.

*This project is supported by DTI Grant IED4/1/5804 and SERC Grant GR/H87629.

1 Introduction

This report describes the motivation, design, collection and analysis of the basic recorded speech database for the first stage of the Cambridge University (Engineering Department (CUED) and Computer Laboratory (CUCL)), and Olivetti Research Limited (ORL) research project on Video Mail Retrieval (Hopper, Sparck Jones & Young 1993). The specification and collection of this database, DATABASE 1, formed task 1 of the overall project plan.

The development of a system to automatically retrieve spoken video mail documents requires an appropriate database of both acoustic training data, for the speech recognition element, and a set of spoken messages, for development and assessment of document retrieval techniques. In order to be suitable this database must fulfil a number of basic requirements (Lamel, Kassel & Seneff 1986) (Sparck Jones 1981). Principally these are that:–

- It must be publicly accessible, hence confidential and personal messages cannot be used.
- It must have reasonable acoustic quality, since automatic speech recognition is not possible on very poor quality speech signals; however it is unrealistic to expect speech recorded in a standard office environment to be of very high quality.
- Sufficient phonetically balanced acoustic training data must be available for the speakers in the database. The messages must have acoustic qualities expected of spontaneous messages encountered in working video mail systems. It would also be useful for the database to contain material for a wide variety of speakers, different ages, sexes, etc, so that the robustness of acoustic models can be assessed. Since the development of the acoustic recognition component of the system is to be an incremental process, beginning with speaker dependent recognition of fixed keywords, the acoustic training data must reflect this. Hence it must contain sufficient training material for this initial baseline system but also be extendable to recognition of untrained keywords in an open speaker environment.
- The overall message set must cover a number of distributed categories or subjects and within these categories a number of different topics. This enables a variety of message requests to be formed, matching documents retrieved and their relevance to the original request assessed. This assessment is important since the effectiveness of the retrieval of messages relevant to the user's need is the overall measure of the system performance. Thus the message set should replicate, as far as possible, the normal information retrieval situation where there are several documents more or less relevant to a request, but not necessarily all matching the set of terms drawn from it well, and there are also non-relevant documents that match well even if most non-relevant documents do not match the search terms.

At the start of the video mail retrieval project such a database did not exist. Various databases had limited coverage of the required attributes but none satisfied them even near sufficiently. For example, the messages collected from normal use of the Olivetti Pandora system have the requisite acoustic qualities of spontaneity but the signal quality

is too poor for speech recognition, some of the messages may not be available for public use (hence all messages would have to be viewed by approved staff to assess this) and the primary user group was very restricted. Other databases designed for development of acoustic wordspotters satisfy the acoustic qualities in terms of training data, signal quality and variety of speakers, however, they do not contain suitable spontaneous messages for assessment of relevance to the user of retrieved documents. Thus in order to be able to proceed with the design of the video mail retrieval system such a database had to be designed, collected and processed.

Since this is a research project it may not be possible to gather all the required data at this point. Hence the data collection design must allow it to be extended as necessary later in the project whilst being complete in itself at this stage. The next section of this report describes the specification of the data collection carried out at this stage. It continues with an account of the actual data collection procedure, the speakers used and the physical collection system. The final section contains a summary of the statistics of the collected database.

For each speaker, Database 1 covers on the one hand read material consisting of isolated words, words in carrier phrases, read sentences, and TIMIT sentences, and on the other freely composed messages. The read material is for acoustic training. The message material can be used both as training material and as test material for both word spotting and retrieval techniques, and in various ways for either purpose. The spoken material has corresponding transcriptions, and we intend the entire database, VMR Database 1, shall be made available to other interested research workers.

2 Specification of Data Collection Task

2.1 Structuring of the Message Set

The structure of the message set to be collected is important if a database of modest size is to be suitable for experimental document retrieval research. The work overhead in message collection and processing is considerable and thus it was only practical to collect such a modest database. This section outlines the principles used to develop the structure of the database.

Clearly it is necessary to get natural messages both in content and speaking style. These messages should be on *topics* within a set of topic *categories*. Since the first stage of the retrieval system development is to operate with a fixed keyword set, each category should have associated with it a set of *keywords* drawn from a limited keyword *vocabulary*. This vocabulary is initially the only set of *terms* to be used for search *queries*.

Topics are individual complex concepts, for example, a topic in the category "wordspotting" might be "the use of whole word models in a fixed vocabulary wordspotting system." It is assumed that there will be a very close relationship between the *content* of a message and its topic. However, a topic is really the essence of a message and may not be expressed within the message in exactly the way specified by the definition of the topic. In general a message can be expected to treat a topic in a finer-grain way than the topic definition. In doing this the message may actually cover more than one specified topic and may even cover more than one specified category.

For simple topic-message specification the messages may be referred to by their spec-

ifying topic. However, this does not imply that messages should actually be labelled with these topics. Rather it is better to consider the message with a summary that could be given to it. This summary may include categories, keywords and/or other words associated with the content of the message.

In particular, very primitive message summaries have been generated for the messages in the database by getting speakers to record a set of *tags* at the time the message is recorded. These have potential utility for fast document searching, like subject fields in email, and can be used for quickly assessing the contents of the database. In the real video mail situation messages have *headers* like ordinary email messages which may be used for searching, e.g. by sender, date, subject line. Database 1 messages do not have such full authentic headers.

As we were collecting messages specially and not drawing on a natural mail community, and also wanted to ensure that both acoustic and retrieval needs were met but from natural messages, we decided to use prompting *scenarios*. These stimulated the speaker to talk on a topic within a category but without constraining them to produce messages on pre-specified topics. A subset of the keyword vocabulary is associated with each category. The scenarios also encouraged the speakers to use keywords for the category, though keywords were not exclusive to categories and could well occur in messages in other categories. It should be noted too that because the scenarios were only prompts, messages for the same scenario could be on quite different individual topics within the same general content area. Speakers were encouraged, but not required, to use category keywords in their messages.

In addition, since the keyword vocabulary is not very large, a set of *otherwords* are specified as an additional prompting vocabulary. A subset of the otherword vocabulary is also associated with each category. These otherwords may be useful, in combination with the fixed keywords, in the introductory work for unrestricted wordspotting later in the project.

In a message prompt the speakers were shown the scenario and both the keywords and otherwords associated with the category to which the scenario belongs.

Finally, it must be emphasised that while it is natural to think, in relation to retrieval, of *requests* indicating topics on which messages are sought and from which particular search queries are derived, the message database has not been constructed with any test requests in mind. The design of requests for retrieval testing is being treated as a separate enterprise.

2.2 Database detail

This section summarises the Database 1 design, showing how we have tried to satisfy our acoustic and retrieval criteria, balancing similarity and difference across messages, within the relatively small set of messages we could realistically expect to gather and transcribe. The database subsumes both a body of messages, for which prompting scenarios are supplied, and utterances of specific phrases, for acoustic training. Both scenarios and phrases are built round sets of particular words, the keywords which constitute the initial search vocabulary for retrieval experiments. The design uses a total keyword vocabulary of 35 words along with 31 related otherwords. The initial design is for the collection of a set of 300 messages satisfying word occurrence requirements both for acoustic word spotting and for realistic term matching in document retrieval. The design is therefore built round

10 broad subject categories providing a 'universe of discourse' environment for naturally–arising specific topics in messages (and future requests), and for a naturalistic distribution of topic information, with similarities and differences, across the message file. Analogous considerations, applying both to the characteristics of their messages and of their speech, determined the number of speakers used and number of messages per speaker.

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

Appendix A contains lists of the message categories, the fixed keywords and the suggested otherwords. Appendix C shows the subsets of fixed keywords and otherwords assigned to each category and the individual scenarios for each category.

The *prompt* for each spontaneous message consisted of the scenario, the keywords for the category and the otherwords for the category. Speakers were asked to favour the use of the listed keywords and otherwords with each message but not at the expense of construction of realistic messages. Also they were not restricted to using the keywords precisely as shown to them but rather to use them freely in their messages, in variant *word forms*. For example, the keyword *mail* could be used in the forms *mailed*, *mails* or *mailing*. The keyword spotter should spot the stem of such words where there is not too much pronounciational variation and count these spots as correct hits on the keyword.

Also, the speakers were not shown a complete list of the keywords used, but rather only those relevant to the category they were currently recording. Hence speakers couldn't avoid, and could naturally use, keywords belonging to other categories.

2.3 Acoustic Training Data

Two types of acoustic training data were also gathered along with the free text messages during the recording sessions.

(i) *keywords in isolation* – 5 examples of isolated utterance of each keyword. These were prompted in a pseudo random order and were taken in groups at the end of each *message set*, as defined by the protocol described later, had been recorded.

(ii) *keywords in carrier phrases* – at least 5 examples of each keyword were embedded in a collection of carrier phrases. A total of 64 carrier phrases were used. These were collected in four sections after each of the four scenario message collection sessions.

The carrier phrases are listed in Appendix B. We also collected a set of 13 *read test sentences*, each keyword occurred twice in this additional set of sentences, in situations similar to those encountered in the carrier phrase sentences. This data can, of course, be used for additional training material. The read test sentences are also listed in Appendix B.

2.3.1 Calibration Data

At the beginning of each recording session speakers spoke the following acoustic calibration sentences :

```
CAL-S1 She had your dark suit in greasy wash water all year.

CAL-S2 Don't ask me to carry an oily rag like that.
```

In addition, at this point of each session a few seconds of "silence" were recorded as a record of the background noise level for the session.

2.4 Sequence of Categories and Scenarios

As we were collecting new messages, not using existing ones, we had to capture some of the distributional effects across message contents and speakers that we believe holds in the real case (though we have not been able to do any serious corpus analysis and have partly argued by analogy with ordinary email), as well as meet the need for acoustic distribution across our vocabulary and speakers. We therefore sought a recording design for messages in categories which would combine coverage of our universe of discourse as a whole with some concentration of messages in content areas.

As mentioned earlier, there are 10 message categories for Database 1. For generality let these be defined as : $C1, C2, \ldots, C10$. Within each category there are 5 scenario prompts, again, for generality let these be defined as : $S1, S2, \ldots, S5$.

2.4.1 Category and Prompt Sequencing

In order to remove sequence effects between categories and between scenarios within categories it is necessary to impose a suitable distributions. These apply both to the order in which the categories are collected for each speaker and to the order in which the scenarios appear within each category for each speaker.

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 in the table 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          1  3  2
2  3  1          3  2  1
3  1  2          2  1  3
```

(from Tague 1981, page 79).

Applying this idea to the sequencing of the data collection task the following sequences of categories and of scenarios are formed.

Category Group Distribution for Speakers The distribution of categories in the groups for the 15 speakers has the following form.

Let the categories given to each speaker be A B C D.

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

Since we have more speakers than categories inevitably some elements will appear more than once in each column i.e. in the same order position in different groups of categories.

Scenario Distribution Since each category is to be used 6 times there must be 6 different sequences of the scenarios for each category. One possibility would be the following.

Let the scenarios within a category be denoted A B C D E.

	A	B	C	D	E
1	S1	S2	S3	S4	S5
2	S2	S5	S4	S3	S1
3	S3	S2	S1	S5	S4
4	S4	S3	S5	S1	S2
5	S5	S1	S2	S4	S3
6	S1	S4	S3	S2	S5

Again since there are more uses of category than scenario some elements must appear more than once in each column.

In the extreme these 6 sequences could be distributed randomly among the 6 occurrences of each category. This might be carrying the decorrelation beyond that needed to claim experimental independence of scenario prompts and would increase the complexity of the collections task considerably. Therefore, we just used the sequences as they appear $1, 2, \ldots, 6$ for each category, as it appears in the collection procedure. This is reasonable since there is no planned or observable correlation between the scenarios of different categories.

Overall, therefore, each speaker had a *category group* of 4 categories, and for each such category recorded 5 messages, one for each scenario, constituting a *message set*; as mentioned earlier, each message set has associated read sentences, constituting a recording *session*, as described later.

2.4.2 Assignment of Speakers to Category Groups

It has been assumed so far (at least implicitly) that the assignment of categories to speakers was random. However, since the task domain here is inevitably fairly artificial it would seem to be a good objective to minimise this artificiality. To this end categories were assigned in such a way so as to direct the groups of scenarios to certain categories of speaker, eg ORL staff to talk about Active Badges and Pandora and CUED staff to talk about Word Spotting. The more general subjects eg Management were used to fill in the gaps in the lists. The rationale behind this strategy was that people are unlikely to send messages on subjects they know nothing about. Using this approach to speaker category assignment, the messages sent should hopefully be more interesting than "Sorry, I don't really know anything about this but, <various half guesses at possible relevant ideas>." Such "guess" messages would have minimal real information content although it might be possible to retrieve them based on their grouped keyword content.

Appendix D shows the assignment of groups categories to each speaker used for the data collection.

2.5 Additional Read TIMIT Training Sentences

In addition to the free messages and the read training data mentioned so far, each speaker recorded 150 *TIMIT sentences* drawn at random from the phonetically rich TIMIT database (Lamel, Kassel & Seneff 1986). This additional data, which is linguistically quite independent of the VMR data so far described, is to be used to train acoustic filler models for the wordspotting, and will also be used to train future subword models. Different random sentence sets were generated for each speaker thus increasing the overall coverage of phonetic variation. This should enable better speaker independent acoustic filler models to be built in the future. The TIMIT training sentence sets were split into 3 equal size sets of 50 sentences.

3 Description of Data Collection Procedure

This section describes the practical details of the data collection. Once again, as with the theoretical database specification, the practical design had to be carefully developed to take into account a number of factors. These requirements and the system developed are discussed in the following subsections.

3.1 Acoustic Signal Considerations

The acoustic quality of the target recognition system for video mail retrieval is specified by the use of the Olivetti Medusa Audio System. This system incorporates a specified desk mounted microphone and custom designed audio preamplifier stage.

For speech recognition systems the current convention is frequently to use very high quality low noise acoustic channels. This is typically implemented by using a high quality headset microphone with an equivalent quality preamplifier stage. Additionally the data is often recorded in a soundproof quiet environment.

It should be clear that the basic quality of the audio signal to be used for video mail retrieval may be considerably inferior to that available for the development of current speech

recognition systems. The reasons for this poorer quality include increased system noise arising from the high preamplifier gain required for a far field microphone, the surrounding acoustic environment arising from both the physical situation and ambient noise conditions, for example office machines, and also continued changes to the background noise caused by random acoustic events. The last category includes events such as surrounding speech, doors being opened and closed, telephones ringing and footsteps.

In order to be better able to take account of these effects, the video mail database was recorded using both microphone types in parallel. Data recorded using the high quality microphone can be used to investigate the optimal available recognition performance. In contrast the data recorded using the desk microphone can be used to investigate performance levels that are anticipated for the Medusa system.

3.2 Recording Situation

The recordings were made in a semi-soundproof room in the ECR Laboratory at CUED. This "Quiet Room" measures five by five metres. The room is closed off by double doors, its windows are double glazed. Fresh air is blown in through a special sound-trapped hole in the wall. Speakers sat on a stable chair in front of a desk. On this desk an unventilated computer monitor was placed. This displayed the sentences and words to be spoken and the message scenario prompts. To preserve the illusion of a video-based interface a video camera was mounted above the monitor and the speaker's image displayed on a monochrome monitor. On the same desk were also placed the mouse and the keyboard of the workstation. The actual workstation box was placed outside the room to keep out ventilation noise. On a desk to the left of the speaker were placed the far-field desk microphone and the preamplifiers for both the close-talking and far-field microphones. The close-talking microphone was attached to a pair of headphones worn by the speaker. The interactive recording process was run by a button clicking interface on the monitor's screen.

It was decided to record in these quiet conditions so that the basic recorded signals would be as noiseless as possible. Investigation of the effect of various noise conditions on recognition performance could be carried out if necessary by mixing separately recorded noise signals in with the clean speech. However, if the speech had been recorded in noisy conditions to start with the noise could not then be removed for comparative tests.

3.3 Recording Equipment

As stated previously, recordings were made using two different microphone systems in parallel on the separate channels of a stereo signal input.

- On the left hand channel, the close-talking head-mounted microphone was recorded. The unit used was a Sennheiser HMD-414 combination headphone–microphone. This is a standard high-quality microphone used for data recordings in the speech community. The output of this microphone was fed into a Symetrix pre-amplifier unit and then into the left channel analogue line input of a Silicon Graphics Iris Indigo workstation. The recordings made on this channel represented the high-quality reference signal.

- On the right hand channel, a desk-mounted far-field microphone was recorded. This microphone was a Canford Audio C100PB Condenser Gooseneck unit, which is used in the standard audio input to the Medusa system. The output of the microphone was fed into a custom-made pre-amplifier unit and then into the right channel of the analogue line input of the Silicon Graphics Iris Indigo workstation. The custom-built pre-amplifier used here was the prototype of the unit used in the audio input of the Medusa system. The recordings made on this channel represented a signal of equivalent quality as would be expected from recordings made using the Medusa system.

The left and right channels of the stereo analogue line input of the Iris Indigo are connected to an internal stereo A/D converter. The signals were sampled in the A/D converter at 16KHz. Recordings were made directly onto the Iris Indigo's hard disk. This was particularly useful since the size of message files meant that considerable delays were introduced if these documents were sent over the network to other disks.

Technical specifications of the recording hardware are contained in Appendix E.

3.4 Interface

The recording system was controlled using an interactive button-operated interface. This was written using the *HGraf* graphical interface tools from the HTK (Hidden Markov model Toolkit) package (Young, Woodland & Byrne 1993). The controls from the interface operated the internal audio functions of the Indigo workstation.

The interface had controls to start and stop recording, pause during recording, play back either left or right channels, go back to review previous recordings from the session, continue after a recording, and quit at the end of a session.

The word or sentence to be spoken or message prompt appeared in a window in the interface. At the end of a recording the waveform of the recording was displayed in another window. Figure 1 shows the control interface used for the data collection system.

Each recording session was constructed in advance and consisted of the sequence of prompts to appear in that session. Each prompt was initiated by a line containing a letter specifying the type of prompt to appear eg read training sentence, scenario prompt. These letters were also combined into the filenames associated with the recorded data. These prompt description letters are listed in Appendix F.

Before starting to record, speakers were given a sheet of paper outlining the purpose of the exercise they were participating in and an overview of the controls in the graphical interface. In addition to this all speakers were supervised through at least their first session to make sure they were familiar with the controls and understood what was required from them.

3.5 Recording Sessions

For each speaker the data recording was split into a total of 8 recording *sessions*. The breakdown of these sessions was as follows:–

- Session 0 : Read Test Sentences, Randomly Distributed Isolated Words.

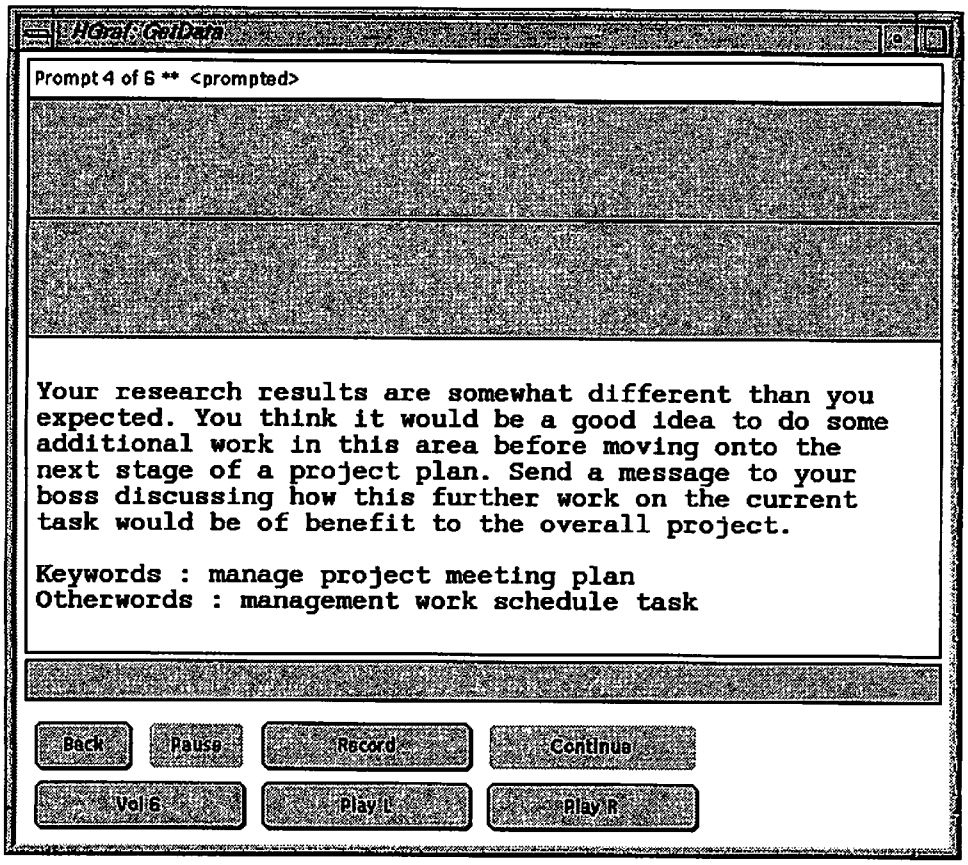

Figure 1: User Interface for Speech Data Capture Software

- Session 1 – 4 : Scenarios SPRx Cy, Read Carrier Phrases 1 – 4, Randomly Distributed Isolated Words
- Session 5 – 7 : Read Phonetically Rich TIMIT Training Sentences

where x represents the speaker number (1–15) and y represents the category number (1–10) for speaker x for the particular session (1–4), as shown in Appendix D.

Before recording started in each session the gain settings on the two input channels were checked independently to ensure they each made good use of the available dynamic range without clipping. The recording for each session began by recording the two calibration sentences listed previously and a few seconds of "silence" as a record of the ambient conditions and channel effects for the session.

Sessions 0 and 5 – 7 typically took in the order of 10 – 15 minutes to record. Sessions 1 – 4 were often rather longer, taking from 20 – 40 minutes each. For these latter sessions the variation in time taken was largely due to the differing length of the messages recorded by different speakers and how long it took them to compose individual messages. All sessions were recorded in sequence but speakers were given a free choice of how many sessions they worked through in each visit. The total number of visits for each speaker varied between 2 and 6.

The prompted messages varied in length from around 20 seconds to the maximum permitted length of around 2 minutes. To encourage speakers to complete messages within the allocated 2 minutes, a warning symbol flashed continuously after about 1 minute 40 seconds.

The amount of message preparation varied widely for the different speakers. Some made brief notes to guide them, others wrote down and read almost exactly what they intended to say and the rest spoke completely without notes. We regard this as representative of the differing levels of preparation which might be found in a user community. Many speakers made effective use of the *pause* button on the recording interface, to produce coherent compact messages.

3.6 Data Storage

For each session the following files were stored:–

- The original sequential prompt file shown to the speaker.
- An information file containing details of the recording session such as the identity of the speaker, the gender of the speaker, the session being recorded, the identity of the operator and the names of the data files stored.
- For each individual read or spontaneous utterance the following were stored:–
  - A "transcription" file of exactly what was shown to the speaker.
  - A speech data file for the close-talking microphone.
  - A speech data file for the far-field microphone.
  - A parametrised file of each speech data file. These files are to be used as the data source for the acoustic wordspotting.
  - For each spontaneous message and tag, a manual transcription of exactly what the speaker said and other acoustic events observed.

The transcription files are considered in more detail in a later section.

Before storage the acoustic speech data files were first losslessly compressed to about 50% of their original size. After this a standard NIST header was added to each file. The details of the NIST header are contained in Appendix F. The header contains a technical description of the stored data.

A standardised file naming scheme was also used. This naming scheme is also shown in Appendix F.

3.7 Speakers

The speakers were recruited from CUED, CUCL and ORL. They included research staff, postgraduate students, managers and secretaries. 11 of the speakers were male and 4 female. Unfortunately, it proved impractical to find more female speakers. For the speaker-dependent acoustic models used for the first wordspotting investigation this imbalance is not important. However, in the longer term the underrepresentation of females in the

training corpus may adversely affect the performance of speaker independent models on females.

For each speaker a record was kept of current age and regional dialect as determined by their primary area of residence between ages 4 and 14 years.

The speakers had a fairly distributed range and level of technical knowledge of the subjects covered by the categories. However, in general they managed to talk successfully about the scenario prompts given to them. It is felt that the range of speakers used probably form a reasonable representation of a group or company which might require the use of a video mail retrieval system. All speakers are familiar with the use of computer workstations. However, some have intimate knowledge of how the retrieval system might work, others are very knowledgeable about abilities of computing systems and perhaps try to help the system by speaking clearly to anticipate recognition problems, while others just recorded messages with no regard for these matters. It should be noted that even where speakers were ignorant of some technical topic they could still generate sensible information-seeking messages, or messages which were both plausible and coherent as discourse even if they were factually incorrect.

But while the distribution of technical proficiency is probably fairly realistic, the speakers used here do not form a natural user community. The speakers were not reacting to situations involving each other but rather often entirely imaginary situations, involving often imaginary or unrelated other users. In fact many of the speakers were completely unknown to each other. Therefore much of the assumed knowledge of an interacting community is not present in these messages. Also none of the messages are replies to other real messages, although some are simulated as replies. This means that it will not be possible to use information such as message sequences to search for these messages.

3.8 Transcription Files

Once collected, the data had to be further processed. For the read speech, it was necessary to verify that the speakers had uttered the actual text correctly. If there was a mismatch the corresponding transcription files were modified to reflect what was actually said. In addition, non-speech events were also indicated, for example loud breaths and tongue clicks. For the spontaneous prompted messages this stage was rather more time consuming. Each spontaneous message had to be fully transcribed, including not only non-speech events but also disfluencies such as partially spoken words, pauses and hesitations such as "um" and "ah." Non-speech events and disfluencies were transcribed in accordance with the Wall Street Journal data collection procedures (Garofoco, Paul & Phillips 1993). Basic punctuation was also added to message transcriptions for ease of reading. The transcribed data is required both for speech training purposes and, at this stage of the project, for retrieval performance testing and bench marking.

3.8.1 Example of prompted message and its transcription

The type of message spoken in response to a scenario prompt and subsequent *message transcription* is probably best demonstrated by example.

In response to scenario OP-SC-3,

```
Send a message asking staff to assess the new output
```

```
routines on the video mail system. Ask them to rank
features such as efficiency of the interface and the
quality of the audio visual output.

Keywords : output score rank assess
Otherwords : list
```

the following message was spoken.

```
[loud_breath] Hello, [pause] as you can see the new video mail system
has been installed [pause] and we are now asking all staff if they can
help us [pause] in assessing the new output routines. [pause] There
are a number of features that we are interested in assessing
[loud_breath] and we are hoping that you can rank these for
us. [pause] For example, the efficiency of the interface, [pause] also
the quality of the audio visual output [pause] and perhaps the
retrieval system. [pause] [loud_breath] We would like to know how you
feel these are working; [pause] perhaps you could give them some score
between zero and ten, [pause] ten being good. [pause] We would also be
happy if you give us some more esoteric answers. [loud_breath] For
example, do you feel comfortable as you use it, [pause] [loud_breath]
is there anything that jars with you? [pause] [loud_breath] Thank you
for your cooperation.
```

With this message the followed tag was recorded.

```
assessing [pause] new [pause] system [pause] routines [pause] factors
```

3.9 Phonetic Transcription & Alignment

An important part of the data processing was the generation of aligned *phonetic transcriptions* needed to initialize HMMs. Therefore an important part of the database is the time-aligned phonetic transcriptions used to initialize our HMM models.

A dictionary based on the Advanced Oxford Learner's Dictionary (70K words) is being developed by CUED Speech Group [1]. This has so far been augmented to a vocabulary of over 100K words and it is hoped that this will be made generally available in the near future. An automatic phonetic transcription tool was created to decompose transcribed text using this dictionary. For this dictionary the standard reduced TIMIT phone set has been augmented with 3 additional vowels specific to British English pronunciation. This augmented phone set is shown in table 1, the 3 additional British English vowels are marked with an asterisk.

Pronunciations for 800 words specific to our database were added to the lexicon to cover out-of-dictionary words as well as disfluencies. The out-of-dictionary words were a combination of proper names, slang, American/British spelling conflicts, compound words, and technical jargon.

[1]The original dictionary was obtained from ftp: black.ox.ac.uk

OTA	TI		OTA	TI		OTA	TI
+			e@	ey		@U	ow
A	aa		eI	ey		oI	oy
&	ae		f	f		p	p
V	ah		g	g		r	r
0	ao		h	hh		s	s
aU	aw		I	ih		S	sh
@	ax		I@	ir	*	t	t
R	axr		i	iy		T	th
aI	ay		dZ	jh		U	uh
b	b		k	k		U@	ur
tS	ch		l	l		u	uw
d	d		m	m		v	v
D	dh		n	n		w	w
e	eh		9	ng		j	y
L	el		N	ng		z	z
3	er	*	O	or	*	Z	zh

Table 1: OTA phonetic alphabet to augmented TIMIT phone mapping

The 35 keywords and 455 keyword variants (plurals, etc.) encountered in the collected data were specially transcribed with keyword-specific phone labels for training keyword-dependent subword and word models. Though the keyword-dependent phone labels appear in the phonetic transcriptions, they may be easily filtered out using a simple regular expression leaving the "raw" phone labels. The full list of keyword-dependent phone labels is contained in Appendix G.

A phonetic transcription has been created for each utterance in the database, using the enlarged lexicon and the text transcription. The phonetic transcription has been automatically aligned with each utterance, using the Viterbi algorithm and American English TIMIT monophone models. Although the alignments appear perfectly satisfactory, when appropriate British English monophones have been trained, they will be used to realign the data.

4 Overview of the Database

This section shows the basic statistics computed for the Database 1 material. Table 2 summarises these statistics for individual speakers and shows the overall average figures for the read and free message material.

It is clear from the table that the length of speakers' messages varies very widely. However, this is almost certainly a characteristic of a real message database. The varying length of message introduces some important issues which will need to be addressed in message scoring for retrieval. All speakers produced a reasonable number of keywords in their messages, although the average number is obviously highly correlated with message

ID	Read Data (min)			Message Data (min)		Test Keywords	Keywords per minute
	"i"	"r"	"z"	"p"	"t"		
1	3.31	5.10	9.06	20.46	1.86	248	12.12
2	4.62	5.92	11.54	18.99	1.79	138	7.27
3	4.89	6.79	11.43	17.52	2.32	186	10.61
4	3.38	6.00	9.71	9.20	3.24	106	11.53
5	4.63	6.45	10.24	7.21	2.21	83	11.50
6	3.28	5.63	9.60	28.00	2.83	277	9.89
7	1.55	3.22	9.64	12.95	1.95	142	10.97
8	4.31	5.88	8.25	25.58	2.44	275	10.75
9	5.27	6.83	8.95	32.88	3.76	246	7.48
10	4.06	6.83	8.59	18.82	1.54	112	5.95
11	3.40	5.31	9.20	23.32	2.03	239	10.25
12	3.81	6.59	8.84	14.52	3.74	130	8.95
13	6.24	7.39	13.59	12.15	5.45	103	8.48
14	3.08	5.41	6.93	10.32	0.96	121	11.72
15	5.51	6.33	10.18	9.27	1.34	128	13.81
Totals	62.64	91.59	145.76	261.20	37.46	2534	
Corpus Totals	Train: 299.99 (5 hr 0 min)			Test: 298.66 (4 hr 59 min)			

Table 2: VMR Speech Corpus Statistics

length.

The total amount of read and message data is nearly identical, but this is obviously a fluke and is not important to the quality of the database.

5 Concluding Remarks

We believe that the 5 hours of spontaneous speech gathered in this data collection form a sufficient database for useful experiments on wordspotting. It is of a comparable size to other databases and very thoroughly transcribed. The information content observed in the messages should also make the database useful for initial information retrieval experimentation as well. However, 300 messages is both too small a set for statistically reliable results and a very small fraction of the number of documents found in a typical information retrieval database and techniques required for larger databases cannot be adequately tested on this small speech document set. We therefore expect to have to develop fuller databases, though not of this expensive and elaborate kind, for later testing.

Quite apart from these needs, for instance, the overall objective of this project is of course *video* mail retrieval, and since this database contains only audio recordings none of the visual prompts that might be associated with the visual content of video messages are present. Users of video mail express their ability to interpret a number of useful features from the visual images. If possible, the assessment of the utility of this feature for video document retrieval will be incorporated into the analysis of the system using real Medusa

video mail messages later in the project.

References

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

[2] Lamel, L.F., Kassel, H.K. and Seneff, S. *Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus,* Proceedings of the DARPA Speech Recognition Workshop, 1986, 26–32.

[3] Sparck Jones, K. *Information Retrieval Experiment*, Butterworths, 1981.

[4] Tague, J.M. "The pragmatics of information retrieval experimentation" in *Information Retrieval Experiment* (Ed Sparck Jones), Butterworths, 1981, 59–102.

[5] Young, S.J., Woodland, P.C. and Byrne W.J. *HTK: Hidden Markov Model Toolkit V1.5,* Cambridge University Engineering Department Speech Group & Entropic Research Laboratories, 1993.

[6] Garfoco, J., Paul, D. and Phillips, M. *CSR WSJ0 Detailed Orthographic Transcription (.dot) Specification,* (ftp: jaguar.ncsl.nist.gov /csr/csr-dot-spec.doc), 1993.

A Keyword, Otherword and Category Definition

The following categories were defined for message collection:-

```
Spotting
Document
Output
Retrieval
Windows
Management
Badge
Pandora
Schedule
Equipment
```

There is also a special additional category `Word Processing`, intended for debugging and experimental use, for which e.g. additional scenarios with special properties may be supplied as needed.

Full list of keywords for the data collection :-

```
 1 : active
 2 : assess
 3 : badge
 4 : camera
 5 : date
 6 : display
 7 : document
 8 : find
 9 : indigo
10 : interface
11 : keyword
12 : locate
13 : location
14 : mail
15 : manage
16 : meeting
17 : message
18 : microphone
19 : network
20 : output
21 : Pandora
22 : plan
23 : project
24 : rank
25 : retrieve
26 : score
27 : search
```

```
 28 : sensor
 29 : spotting
 30 : staff
 31 : time
 32 : video
 33 : windows
 34 : word
 35 : workstation

Total : 35 keywords
```

Full list of otherwords for the data collection :-

```
  1 : active-badge
  2 : browse
  3 : deadline
  4 : disc
  5 : display
  6 : effort
  7 : file
  8 : get
  9 : graphics
 10 : hit
 11 : item
 12 : keyboard
 13 : list
 14 : listen
 15 : management
 16 : match
 17 : microphone
 18 : noise
 19 : retrieval
 20 : schedule
 21 : score
 22 : screen
 23 : server
 24 : speak
 25 : system
 26 : task
 27 : unix
 28 : video_mail
 29 : work
 30 : xab
 31 : X-windows

Total : 31 otherwords
```

B Read Training and Test Data

B.1 Carrier Phrases Section 1

CP-S1-1 The camera locations are good for spotting mistakes.

CP-S1-2 Word spotting is basically finding the location of important keywords.

CP-S1-3 Two microphones are connected to each Indigo workstation.

CP-S1-4 I like your window-based interface. The display looks good.

CP-S1-5 What dates are you free? It won't take much time.

CP-S1-6 Do we just rank the output by score?

CP-S1-7 Careful planning and management helps most projects.

CP-S1-8 Word spotting systems need a good search strategy to find the keywords.

CP-S1-9 The score and rank of each message require careful assessment.

CP-S1-10 How many windows can the interface display?

CP-S1-11 A Pandora system is a workstation, a video camera, and a network interface.

CP-S1-12 Time is running out. Meeting the project date depends on available staff.

CP-S1-13 Can I retrieve my video mail if my workstation has no camera?

CP-S1-14 Of all the dates, that's the worst date. Perhaps another time?

CP-S1-15 Spotting active windows is easy; they're located at the top of the display.

CP-S1-16 Pandora video quality depends more on the network than the workstation or the camera.

B.2 Carrier Phrases Section 2

CP-S2-1 I think the management plan is hard to assess.

CP-S2-2 The new network should have better video output quality than Pandora.

CP-S2-3 Planning a project usually means several management

meetings.

CP-S2-4 Can I use the microphone over the network? My workstation is not an Indigo.

CP-S2-5 Finding keyword locations can be difficult, especially when spotting short words.

CP-S2-6 Your manager will explain how to mail a message.

CP-S2-7 The Pandora retrieval system displays the scores.

CP-S2-8 The spotting system retrieves only active keywords.

CP-S2-9 Not all the Indigos have a microphone in that location.

CP-S2-10 Let's plan an agenda for the management meeting. Which project is first?

CP-S2-11 Where's the display located?

CP-S2-12 Will unplugging a sensor stop the active badge network?

CP-S2-13 Why can't it locate the keyword? Did it retrieve the document rank?

CP-S2-14 Do you think windows are the best display interface?

CP-S2-15 A "hit" is when the search finds a keyword.
A "miss" is when a search doesn't find a keyword.

CP-S2-16 Other sensors are located on the Indigos. This gives a more precise location.

B.3 Carrier Phrases Section 3

CP-S3-1 Badge sensors can be used to locate staff in different locations.

CP-S3-2 Most managers plan project schedules.

CP-S3-3 A mail document may contain only a simple message.

CP-S3-4 The active badge network has several sensor types.

CP-S3-5 Most sensors are located on a wall or door.

CP-S3-6 The display interface uses windows.

CP-S3-7 I want to retrieve a particular message from a number of documents.

CP-S3-8 How do I restart the Indigo? The network and microphone are down.

CP-S3-9 To quickly assess the outputs, rank them by score.

CP-S3-10 How do I retrieve my mail messages?

CP-S3-11 The Indigo is good for this project since it has both microphone inputs and network connections.

CP-S3-12 Others can locate you only if you wear your active badge.

CP-S3-13 Who is the manager on this project?

CP-S3-14 How do you actively assess the rank?

CP-S3-15 Each mail document contains a message header.

CP-S3-16 The sensor location must be high enough to reliably locate an active badge.

B.4 Carrier Phrases Section 4

CP-S4-1 We had better assess the score quality before we rank them.

CP-S4-2 Just assess each output to determine its score.

CP-S4-3 Video mail can be sent from any Pandora workstation with a video camera.

CP-S4-4 The Pandora workstation records video mail using the microphone and camera.

CP-S4-5 Data retrieval is more than just a search.

CP-S4-6 I like the document display, but how do you find the search output?

CP-S4-7 Some searches take too long to find documents.

CP-S4-8 Tell the staff about the meeting date and time.

CP-S4-9 Active badges help staff to plan their time.

CP-S4-10 I set the meeting date last week. Where's the rest of the staff?

CP-S4-11 There's an active badge sensor in each room.

CP-S4-12 That plan is out of date. Time for a staff meeting!

CP-S4-13 Planning staffing requirements always takes time.

CP-S4-14 Did it find the keyword? Was the message retrieved?

CP-S4-15 What date is the next meeting? Staffing might be a problem if the time is too early.

CP-S4-16 The interface is hard to assess. Light from the window glares off the output display.

B.5 Read Test Sentences

TS-1 This project concerns how to find video mail messages in a document archive.

TS-2 Meeting the project plan will require collecting many sample messages.

TS-3 Messages are recorded on a Pandora workstation. Each Pandora on the network has a video camera as well as a microphone.

TS-4 An active microphone is also located on the desk. Video from the camera is displayed for the user.

TS-5 Finding messages is a two-stage process. First, word spotting techniques are used to search and locate keywords in each video mail document.

TS-6 Second, information retrieval methods are used to score and rank each document on the basis of the keyword spotting output.

TS-7 A window-based interface will output the retrieved documents, ranked by score, for the user to assess.

TS-8 The management plan of the video mail project displays the necessary time and staff requirements, and scheduled date of completion.

TS-9 Management meetings will assess whether the project is proceeding within the allotted time window.

TS-10 Each staff member has an active badge which registers location, date and time.

TS-11 The badges are detected by a network of sensors found at locations throughout the building.

TS-12 Staff members find that indigo workstations are best for message search and retrieval.

TS-13 An indigo sensor interface actively displays the location and time of badge spottings.

C Category Keyword and Otherword Specification and Scenario Prompts

For each category listed in appendix A, the following are listed :-

- fixed keywords
- suggested otherwords
- the scenario prompts for speaker messages

```
CAT-1 SPOTTING (SP)

   KEYWORDS (SP-KW)
     spotting  word  search  keyword  find

   OTHERWORDS (SP-OW)
     match  hit  score  noise

   SCENARIOS

      SP-SC-1 Acoustic word spotting is widely used to search a file
              to find given important keywords. Suggest some
              situations where using a keyword to find important
              words might be useful.

      SP-SC-2 Describe some sources of noise which might make it
              difficult to find the keywords in a word spotting
              search.

      SP-SC-3 Send a message in response to an enquiry into how using
              word spotting to find keywords may be more useful than
              attempting to find every word in an utterance

      SP-SC-4 Form a message to outline how criteria might be combined
              to maximise the chance of a hit (correct match), in a
              word spotting system. For example, the comparison with
              other acoustic speech information and the ambient
              background noise.

      SP-SC-5 Outline some of the potential implications of very poor
              match performance in a word spotting system. For
              example, missing important words and finding words that
              aren't there.

CAT-2 DOCUMENT (DO)

   KEYWORDS (DO-KW)
     document  message  mail

   OTHERWORDS (DO-OW)
     item  file video_mail
```

SCENARIOS

DO-SC-1 Many mail documents contain useful long term information. However, often a message very quickly becomes out of date. Send am message to staff containing a policy outline about the importance of removing old documents once their messages are out of date.

DO-SC-2 The sending of junk mail is an increasing problem on computer networks. The useful message content of these documents is often nil and they are frequently very long. Imagine having to deal with such junk video mail. Compose a message to be sent to mail servers requesting that unsolicited video mail message documents are not forwarded to you.

DO-SC-3 Mail messages are frequently composed of a number of items of information in a single document. Generate a message asking staff to send single item messages since cataloguing and retrieval of multi-item documents is very difficult.

DO-SC-4 A method of accessing a mail document easily is to assign a tag of a few words to describe the message content. Send a message to staff stressing the importance of the tag in retrieval of video mail documents. Also that words in the tags should be spoken clearly to assist acoustic recognition.

DO-SC-5 Discuss the uses of using computer mail to send documents containing messages which could just as easily be sent by phoning someone or even shouting across the room. Example, permanent record in a file, don't have to look for them to pass the message.

CAT-3 OUTPUT (OP)

KEYWORDS (OP-KW)
output score rank assess

OTHERWORDS (OP-OW)
list

SCENARIOS

OP-SC-1 Send a message containing your comments about the details given at the output of an information retrieval search of a video mail archive. What non acoustic keys could be included to help find the message you're looking for ?

OP-SC-2 Consider factors which might be taken into account to calculate the score to rank the output of a video mail retrieval search.

OP-SC-3 Send a message asking staff to assess the new output routines on the video mail system. Ask them to rank features such as efficiency of the interface and the quality of the audio visual output.

OP-SC-4 Do you think use of video images in the output of a retrieval search would help to assess whether you've found the message you were looking for. You may or may not have seen it before and you may not even know the speaker.

OP-SC-5 Send a message reminding staff to enter all the appropriate tag and identification data for each video mail message. Explain that the algorithms use to score and rank the message retrieved in response to an enquiry become inefficient otherwise and that this means that long lists of messages are needing to be retrieved at present.

CAT-4 RETRIEVAL (RE)

KEYWORDS (RE-KW)
retrieve search find display

OTHERWORDS (RE-OW)
retrieval browse listen get

SCENARIOS

RE-SC-1 The automatic retrieval of information requires an efficient search scheme to find relevant documents. Form a message consider the possible information about an item which if stored could be used to help search for it eg topic, speaker, date.

RE-SC-2 Once a retrieval system has found what it regards as the relevant messages, how might the speaker find those he or she intended to get? After all, one could look at them all but this would be rather inefficient.

RE-SC-3 Message retrieval can take several iterative passes: each one can search for particular keys, eg keywords. Changing the keys to refine the search after each pass can lead to more selective and better quality output. Generate a message considering how the search might be refined at each pass.

RE-SC-4 What do you think would be a good way of browsing the output of a document retrieval system. This could

involve a display, listening or both in an attempt to find documents meeting your needs.

RE-SC-5 The amount and quality of the information given to an retrieval system in the request to search will affect not only the chance of finding the message(s) you like but also the number of additional messages. Generate a message considering how you might go about forming efficient queries.

CAT-5 WINDOWS (WI)

KEYWORDS (WI-KW)
windows display interface

OTHERWORDS (WI-OW)
X-windows screen graphics server

SCENARIOS

WI-SC-1 Describe some of the advantages of windows as a screen interface on modern workstations as opposed, for example, to single screen systems with poor graphics.

WI-SC-2 Outline the features which you feel are important for a windows display, for example, an editor, system loads display and icons.

WI-SC-3 The windows environment can be extended to include display of video mail messages and entertainment such as TV channels. What effect do you think the introduction of these new types of windows with their facilities would have on the workplace, greater productivity from a better interface or all day TV watching ?

WI-SC-4 I'm told that X-windows can be really attractive if it is set up with a well defined set of menus. Can you suggest how elements should be placed in menus to improve my screen setup ? I'm worried that it might become more confusing if there are too many menus.

WI-SC-5 A colleague is uncertain as to why your new graphics package doesn't work in the standard windows environment. However, they haven't explained what goes wrong in sufficient detail. Send a message asking what is wrong with suggestions of what might be seen on the screen and possible causes.

CAT-6 MANAGEMENT (MA)

KEYWORDS (MA-KW)
manage project meeting plan

OTHERWORDS (MA-OW)
management work schedule task

SCENARIOS

MA-SC-1 Your current project is lagging behind the 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.

MA-SC-2 Your research results are somewhat different than you expected. You think it would be a good idea to do some additional work in this area before moving onto the next stage of a project plan. Send a message to your boss discussing how this further work on the current task would be of benefit to the overall project.

MA-SC-3 Our clients have failed to make the agreed delivery date. If we don't reschedule the related work we will waste a lot of time. Send a message asking for suggestions of how to modify the project task schedule so that we can still meet the overall project deadline.

MA-SC-4 You are falling badly behind the schedule for your current task. Send a message to the project management team pointing this out and expressing your concern about the effect of this on the overall plan.

MA-SC-5 Someone has asked you to become involved with the management of a new project, however, the project plan they have sent you is rather poor. Send a message expressing what you think of the plan, how a suitable plan might be developed and your overall feelings about being involved.

CAT-7 BADGE (BA)

KEYWORDS (BA-KW)
badge active sensor locate location

OTHERWORDS (BA-OW)
system xab active-badge

SCENARIOS

BA-SC-1 Send a message describing the basic ides of an active badge including such details as the need for a sensor and a workstation poller.

BA-SC-2 Send a message in response to an enquiry concerning the efficiency of sensor location and where you might be

able to connect your active badge network to.

BA-SC-3 The active badge system is to be integrated with the Pandora multimedia workstations. Send a message in response to an enquiry into the problems involved and how solutions might be scheduled. Possible problems are with the availability of sensors, software compatibility and opening of xab windows on the video output of the Pandora.

BA-SC-4 Send a message to the maintainer of your company active badge system describing what you think is a fault with sensor because the system seems to find you everywhere except your office. Suggest suitable times that would be convenient to come and check it.

BA-SC-5 The current active badges have only limited interactive facilities, the user can force a beacon and the badge can beep. Consider some additional features which might be added, bearing in mind the fact that there are fixed sensors in each location.

CAT-8 PANDORA (PA)

KEYWORDS (PA-KW)
Pandora workstation camera video mail

OTHERWORDS (PA-OW)
speak microphone

SCENARIOS

PA-SC-1 The Pandora is a high quality multimedia workstation. Discuss.

PA-SC-2 Send a message objecting to the lack of privacy arising from the fact that the camera of the Pandora workstation is permanently active.

PA-SC-3 Compose a message discussing the features which already find or would find attractive in a multimedia workstation such as Pandora. Examples are video mail. cd music, TV, video phone plus other audio/visual features you can think of.

PA-SC-4 Send a video mail message asking Pandora users to archive their mail messages carefully since retrieval is made difficult if the file names are not clear. Because of this often many messages need to be searched to find the right one.

PA-SC-5 Send a message asking staff to refrain from using the multimedia facilities on Pandora today since an

important visitor is expected, ie stop watching TV and listening to CDs. Also ask them to position there cameras to point at them at work so that the video phone etc can be demonstrated with calls to staff at random.

CAT-9 SCHEDULE (SC)

KEYWORDS (SC-KW)
time date meeting staff

OTHERWORDS (SC-OW)
task deadline effort

SCENARIOS

SC-SC-1 Send a message informing the staff of the time and date of the next progress meeting for your current project. (Make up something realistic).

SC-SC-2 Send a message enquiring into the deadline for the completion of your present task.

SC-SC-3 Send a message enquiring into the effort that will be required by different members of staff to meet a required deadline. Specify in the message the date and time by which the task must have been completed.

SC-SC-4 Send a message suggesting times and dates of the staff meetings. Also specify the deadlines for corresponding progress reports which must be submitted in advance of these meetings.

SC-SC-5 Send a message to your boss in reply to his message pointing out that the company is about to miss the project deadline and have to make the contracted late delivery payments. In your message suggest a date by which the work will probably be completed and suggest a time and date for a meeting to sort this out.

CAT-10 EQUIPMENT (EQ)

KEYWORDS (EQ-KW)
indigo workstation microphone network

OTHERWORDS (EQ-OW)
keyboard unix display disc

SCENARIOS

EQ-SC-1 Send a message asking someone to come to look at the microphone input to your indigo since it doesn't seem to be working. Suggest some times when you will be

available to demonstrate the problem.

EQ-SC-2 Disc access on my indigo seems to be very slow. Do you think there might be a problem with my unix setup, the network or the disc reading ?

EQ-SC-3 Entry of acoustic data via the microphone and keyboard control don't seem to be synchronised on my indigo. Can you suggest anything I could do to sort this out or how I mught go about getting hold of some information ?

EQ-SC-4 Send a message asking people to clear their files off the disc on your indigo since you need fast local disk access for your next job which involves acoustic data collection via the microphone input.

EQ-SC-5 Send a message to your research group requesting them to refrain from running jobs on your indigo since you have to get some results in a hurry. Tell them that if they don't do this themselves you'll use unix to do it for them.

CAT-DEBUG WORD PROCESSING (WP)

KEYWORDS (WP-KW)
latex spellcheck document edit

OTHERWORDS (WP-OW)
postscript emacs format paper

PHRASES

WP-PH-1 When using latex it can be rather tiresome to spellcheck a document although it is easy to edit it.

WP-PH-2 If you edit a document it is important to spellcheck it before using latex.

WP-PH-3 A spellcheck system is essential in wordprocessing if each document is not be errorful and needing to go through the edit stage again.

WP-PH-4 Most scientific journals will not accept a document in latex format even if they edit it to another format.

WP-PH-5 Stripping latex commands from a document is not straightforward, it is often necessary to edit the document by hand and spellcheck it again.

WO-PH-6 You can spellcheck a latex document using ispell.

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 ?

D Assignment of Groups of Categories to Each Speaker

Taking into account the comments about assignment of categories to speakers who would hopefully be able to speak with some knowledge in section 2.4.1, the following category assignments were made. These were done 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 speaker groups. Despite this apparent restriction the category distribution is still observed to be quite broad.

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

For the individual speakers the following subject groups were produced:

```
SPR1  : C1  spotting
        C2  output
        C3  document
        C4  management

SPR2  : C5  badge
        C6  Pandora
        C7  retrieval
        C8  schedule

SPR3  : C9  windows
        C10 equipment
        C1  spotting
        C2  output

SPR4  : C3  document
        C4  management
        C5  badge
        C6  Pandora

SPR5  : C7  retrieval
        C8  schedule
        C9  windows
        C10 equipment

SPR6  : C1  spotting
        C3  document
        C5  badge
        C7  retrieval
```

```
SPR7  : C9  windows
        C2  output
        C4  management
        C6  Pandora

SPR8  : C8  schedule
        C10 equipment
        C1  spotting
        C3  document

SPR9  : C5  badge
        C7  retrieval
        C9  windows
        C2  output

SPR10 : C4  management
        C6  Pandora
        C8  schedule
        C10 equipment

SPR11 : C2  output
        C5  badge
        C7  retrieval
        C1  spotting

SPR12 : C6  Pandora
        C3  document
        C8  schedule
        C9  windows

SPR13 : C10 equipment
        C4  management
        C2  output
        C5  badge

SPR14 : C7  retrieval
        C1  spotting
        C6  Pandora
        C3  document

SPR15 : C8  schedule
        C9  windows
        C10 equipment
        C4  management
```

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

Using the randomisation of scenario order within each category, each instance of a category shown here represents the delivery of the scenarios in a different order in accordance with the random scenario distribution described previously.

E Specifications for Data Collection Audio Equipment

The Far-field Desk Microphone: *Canford C100PB Condenser Gooseneck*
(Official Technical Specifications)

Output Impedance :	$2k\Omega \pm 20\%@1kHz$
Sensitivity :	$-64dBV \pm 3dB(0dB - 1V/\mu$ bar @$1kHz)$
Polar Response Cardioid :	Front to back rejection $10dB$ approx. $(1kHz)$

The Head-mounted Close-talking Microphone: *Sennheiser HMD 414-6*
(Official Technical Specifications)

Frequency Response :	$50Hz - 12kHz$
Mode of Operation :	Pressure gradient transducer for close talking
Directional Characteristic :	super-cardioid
Rejection at 120° and 1000Hz :	$20dB - 2dB$
Impedance at $1000Hz$:	200Ω
Sensitivity :	$1\mu V/50mG = 1\mu V/5\mu T$

The Head-mounted Microphone Pre-amplifier: *Symetrix SX202 Dual Mic Preamp*
(Official Technical Specifications)

Frequency Response :	$20Hz - 20kHz, +0dB, -1dB$
SNR :	$95dB(-50dBV, 150\Omega)$
Max. Gain / Min. Gain :	$60/20dB$

Silicon Graphics IRIS Indigo **: Stereo Line-Level Analog Input**
(Official Technical Specifications)

Nominal Input Impedance :	$5k\Omega$
Input Signal :	
Max. Amplitude :	$10Vpp$
Min. Level :	$1Vpp$ (for full-scale input)

Silicon Graphics IRIS Indigo **: A/D Converter**
(Official Technical Specifications)

Resolution :	Stereo 16-bit
Modulation :	delta-sigma
Used Sampling Rate :	16kHz
Oversampling :	64x
Official SNR at $48kHz$:	$> 80dB(20Hz - 20kHz)$

F Speech Data File Format Specifications

The speech data was encoded using the following file format.

```
Encoding: 16 kHz 16-bit pcm
Header: NIST SPHERE format (1K ascii)

NIST_1A
   1024
speaking_mode -s11 read-common
recording_site -s4 CUED
recording_environment -s14 ECR_Quiet_Room
database_id -s8 VMR_set0
channel_count -i 1
sample_count -i <var>
sample_max -i  <var>
sample_min -i  <var>
sample_rate -i 16000
sample_n_bytes -i 2
sample_byte_format -s2 10
sample_sig_bits -i 16
file_id -s8 <var>
talker_id -s3 <var>
info_file -s8 <var>
talker_gender -s  < male | female >
microphone_used -s18 < Sennheiser_HMD_414 | C100PB_Condenser_desk >
zero_mean -s3 yes
end_head
```

The mode of recording for a recording prompt was marked by a letter in the first line of the prompt (not shown to the user). The mode letter definitions were as follows:–

- "r" – for a read training sentence for the specified keywords.
- "i" – for an example of one of the isolated keyword utterances.
- "z" – for one of the phonetically rich TIMIT sentences.
- "p" – for a message prompt.
- "t" – for a tag summary of a prompted message.
- "c" – for a calibration sentence.
- "b" – for a recording of ambient background conditions

The file_id, which also serves as the filename root, is an 8–character string composed of the following fields:–

```
   XnnsYmmm   where

 X = B | G | M | F:    Talker gender and microphone used
                       M, F: male/female head microphone
                       B, G: male/female desk microphone

   nn:                 2-digit unique speaker id

   s:                  1-digit unique session id

Y = b | c | i | p | r | t | z :
                       1-character mode code

   mmm:                3-digit unique utterance id
```

For example, `M513z024` is an utterance of a TIMIT read sentence, spoken in session 3 by talker 51 (who is male), recorded with the close–talking microphone.

G Keyword–specific phone labels

ACTIVE	ae_01 k_01 t_01 ih_01 v_01
ASSESS	ax_02 s_02 eh_02 s_02A
BADGE	b_03 ae_03 jh_03
CAMERA	k_04 ae_04 m_04 ax_04 r_04 ax_04A
DATE	d_05 ey_05 t_05
DISPLAY	d_06 ih_06 s_06 p_06 l_06 ey_06
DOCUMENT	d_07 ao_07 k_07 y_07 uh_07 m_07 eh_07 n_07 t_07
FIND	f_08 ay_08 n_08 d_08
INDIGO	ih_09 n_09 d_09 ih_09A g_09 ow_09
INTERFACE	ih_10 n_10 t_10 ax_10 f_10 ey_10 s_10
KEYWORD	k_11 iy_11 w_11 er_11 d_11
LOCATE	l_12 ow_12 k_12 ey_12 t_12
LOCATION	l_13 ow_13 k_13 ey_13 sh_13 n_13
MAIL	m_14 ey_14 l_14
MANAGE	m_15 ae_15 n_15 ih_15 jh_15
MEETING	m_16 iy_16 t_16 ih_16 ng_16
MESSAGE	m_17 eh_17 s_17 ih_17 jh_17
MICROPHONE	m_18 ay_18 k_18 r_18 ax_18 f_18 ow_18 n_18
NETWORK	n_19 eh_19 t_19 w_19 er_19 k_19
OUTPUT	aw_20 t_20 p_20 uh_20 t_20A
PANDORA	p_21 ae_21 n_21 d_21 or_21 r_21 ax_21
PLAN	p_22 l_22 ae_22 n_22
PROJECT	p_23 r_23 ao_23 jh_23 eh_23 k_23 t_23
RANK	r_24 ae_24 ng_24 k_24
RETRIEVE	r_25 ih_25 t_25 r_25A iy_25 v_25
SCORE	s_26 k_26 or_26 axr_26
SEARCH	s_27 er_27 ch_27
SENSOR	s_28 eh_28 n_28 s_28A ax_28 axr_28
SPOTTING	s_29 p_29 ao_29 t_29 ih_29 ng_29
STAFF	s_30 t_30 aa_30 f_30
TIME	t_31 ay_31 m_31
VIDEO	v_32 ih_32 d_32 ir_32 uh_32
WINDOWS	w_33 ih_33 n_33 d_33 ow_33 z_33
WORD	w_34 er_34 d_34
WORKSTATION	w_35 er_35 k_35 s_35 t_35 ey_35 sh_35 n_35