# Video Mail Retrieval by Voice: Towards Intelligent Retrieval and Browsing of Multimedia Documents

J. T. Foote[1] M. G. Brown[2] G. J. F. Jones[1,3] K. Sparck Jones[3] S. J. Young[1]

[1]Cambridge University Engineering Department, Cambridge, CB2 1PZ, UK
[2]Olivetti Research Limited, 24a Trumpington St., Cambridge, CB2 1QA, UK
[3]Cambridge University Computer Laboratory, Cambridge, CB2 3QG, UK

## Abstract

*This paper describes a multimedia document retrieval project at Cambridge University and Olivetti Research Limited (ORL). The project seeks to integrate state-of-the-art text retrieval methods with high-performance word spotting to yield a robust and efficient video mail retrieval system. A specific goal is the development of a practical retrieval system to work with Medusa, a high-bandwidth multimedia environment in daily use at Olivetti Research Limited. This paper describes the project background, and presents recent results showing audio retrieval performance close to that of text. First steps towards an intelligent user interface are presented, including a "video message browser" that represents a mail message as a static image that can be reviewed at a glance. The browser allows interesting portions of the message to be found, selected, and played back at the user's convenience.*

## 1 Introduction

The last few years has seen an increasing use of multimedia applications, including video conferencing and video and audio mail. Using these facilities can create large archives of video material, which poses a significant storage, retrieval, and access problem. Users are unable to find stored messages because, unlike text, there are no simple ways to search for a particular reference. Even when a message is found, it may be difficult to judge the usefulness of the message without replaying it in its entirety, which is inefficient and time-consuming.

The Video Mail Retrieval (VMR) project is addressing this problem by developing a system to retrieve and browse stored video messages by intelligently interpreting the audio stream. The project seeks to integrate state-of-the-art text retrieval methods with high-performance word spotting, and to present the resultant multi-modal information in an intelligent user interface. A specific goal of the project is to develop a useful retrieval application to work in the Medusa multimedia environment developed at Olivetti Research Ltd in Cambridge [1].

In the simplest form of message retrieval, a user specifies a single search keyword and word spotting techniques locate its occurrences in the audio soundtrack. A more robust system uses multiple search keys, both to minimize the effect of spotting errors and to refine the list of retrieved messages. Thus, the topic specification and search strategies developed for conventional text-based information retrieval (IR) must be adapted to this new environment [2]. Although later stages of the project will investigate open-keyword and open-user sets, the initial stage, and the work described here, focuses on a fixed, *a-priori* known set of search keywords and users.

## 2 The Video Mail Retrieval System

A prototype Video Mail Retrieval application has been developed that integrates keyword spotting and information retrieval. Figure 1 shows a block diagram of the VMR system, which operates as follows. The audio soundtrack of each message (whether from an existing archive or received as new mail) is passed

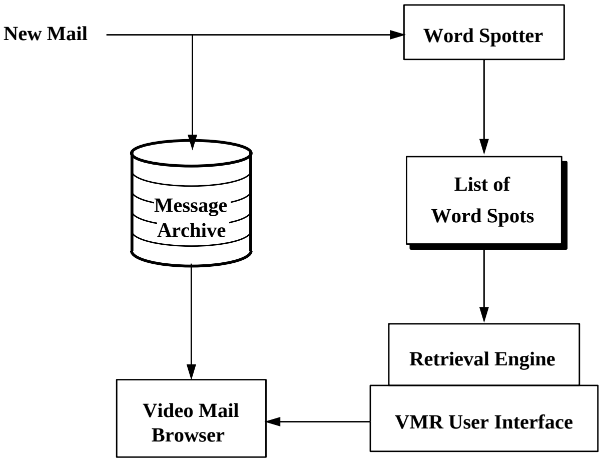


Figure 1: Block diagram of VMR system.

to the acoustic word spotter. This computes a sequence of putative keyword hits, which is added to an index containing putative hits for all messages, along with mail header information and pointers to the message data (for playback). Because the keyword set is fixed and known in advance of requests, the computationally-intensive word spotting is done as messages are added to the archive. Thus retrieval time is nearly instantaneous for the user application.

## 2.1 Keyword Spotting

Automatically detecting words or phrases in unconstrained speech is termed "keyword spotting;" this technology is the foundation of the VMR system [3]. The best-performing keyword spotters are based on the same hidden Markov model (HMM) methods used in successful continuous-speech recognition [4]. A hidden Markov model is a statistical representation of a speech event like a word; model parameters are typically trained on a large corpus of labelled speech data. Given a trained set of HMMs, there exists an efficient algorithm for finding the most likely model sequence (the recognized words), given unknown speech data.

In a real-world system, the message archive of Figure 1 would typically contain video mail messages intended for a user or group. For the initial development of the VMR system, it was necessary to create a test archive of messages with known audio and information characteristics. The VMR1 message corpus is a structured collection of audio training data and information-bearing audio messages. Because of storage limitations at the time of collection (now overcome), only audio was collected, so the system described here retrieves audio-only messages. Ten "categories" were chosen to reflect the anticipated messages of actual users, including, for example, "management" and "equipment." A fixed set of 35 keywords were chosen to cover the ten categories; thus the "management" category includes the keywords "staff," "time," and "meeting." Fifteen subjects each provided about 45 minutes of speech for a total of nearly ten hours of recorded data. Half of this data was read speech necessary for acoustic model training; the remaining 5 hours of speech was 300 spontaneous messages used to evaluate both word spotting and information retrieval performance. Speech was recorded in high-fidelity using both a desk and close talking microphone. (A full description of the VMR corpus may be found in [5].)

For hidden Markov model (HMM) training and recognition, the acoustic speech data were parametrised into a spectral representation. Whole-word talker-dependent keyword models and monophone filler models were constructed for each of the 15 talkers, using the HTK tool set [6]. While a full treatment of the keyword spotting may be found elsewhere [3, 7], the basic results are presented in Figure 2. This graph shows that at the moderate false alarm rate of 10 false alarms per keyword per hour, nearly 90% of keywords are correctly spotted.

## 2.2 Information Retrieval

Information retrieval experiments require message *queries* and assessments of message *relevance* to queries. A set of 50 queries was generated from 10 users, in the form of a typed natural-language request. Each user then assessed the relevance of archive messages to queries they produced [8]. In retrieval, a relevance

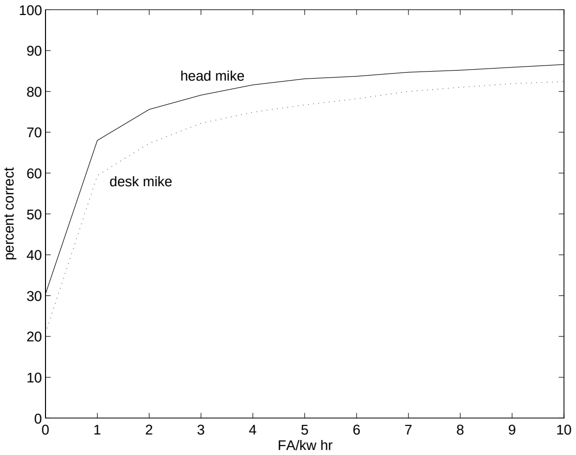


Figure 2: VMR Task Keyword Spotting ROC (Receiver Operating Characteristic)

| | Ave. Prec. | Text Rel. | Phon. Rel. |
|---|---|---|---|
| Text | 0.332 | 100% | — |
| Phonetic | 0.339 | 102.4% | 100% |
| Head | 0.312 | 94.0% | 91.8% |
| Desk | 0.296 | 89.4% | 87.3% |

Table 1: Acoustic retrieval performance (absolute and relative average precision).

score may be computed from a weighted sum of the keywords common to both the message and query [9, 10]. Messages are ranked by score and presented to the user. The automatically-generated scores may be compared with the human-assessed relevances to judge how well the retrieval is working; a common measure is the *average precision*, which is roughly the proportion of retrieved messages that are actually relevant.

### 2.3 Information Retrieval from Word Spotting

Acoustic word spotting is prone to false alarms and missed keywords, so retrieval can be expected to suffer. The extent of the degradation can be measured by comparing the retrieval performance on word spotting results to the retrieval performance on text. An additional problem of word spotting is that unrelated acoustic events will often resemble valid keywords. For example, the last part of "hello Kate" is acoustically quite similar to the keyword "locate."

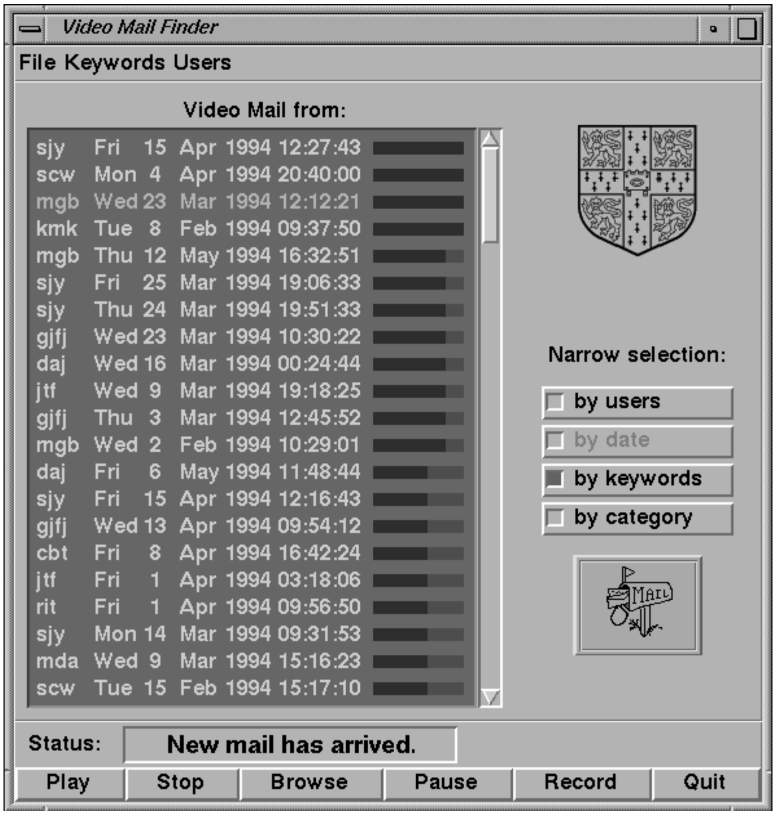
Because even the best acoustic models cannot discriminate between homophones, the output of an ideal word spotter that reports all keyword phone sequences provides a more legitimate standard of comparison than text. This was simulated by scanning the message phonetic transcriptions for sequences that match keyword phone sequences. The word spotter outputs a list of putative keyword hits and associated acoustic scores. Since the current message retrieval scheme uses only the presence/absence of a keyword in a message, the acoustic scores are thresholded such that only hits with a score above the threshold are counted.

The best threshold value represents an optimal trade off between the numbers of true hits and false alarms. Table 1 compares acoustic retrieval performance to the ideal text and phonetic standards, at the *a-posteriori* best thresholds. Though performance with data from the desk microphone is slightly worse, it is encouraging that retrieval performance is degraded only 10–15% from the ideal by using actual word spotter output.

## 3 The Video Mail Retrieval User Interface

At some point, results from both the keyword spotting and information retrieval must be presented to the user. The approach taken for the VMR user interface is the "message list filter." Upon startup, a scrollable list shows all available messages in the user's video mail archive. Using information in the mail message header, various controls let the user "narrow" the list, for example, by displaying only those messages from a particular user or received after a particular time. Unsetting a constraint restores the messages hidden by that constraint; multiple constraints can be active at one time, giving the messages selected by a boolean conjunction of the constraints.

A natural addition to this scheme is to add message attributes that depend on the keywords spoken in the audio portion. Messages can then be ranked by the presence/absence of a keyword or combination of keywords. In operation, the user selects one or more keywords as a search query. The resulting score for each message is computed by the retrieval engine, and the interface then displays a list of messages ranked by score. Scores are represented by bar graphs, as in Figure 2; messages with identical scores are ranked



Video Mail from:

| | | | | | |
|---|---|---|---|---|---|
| sjy | Fri | 15 | Apr | 1994 | 12:27:43 |
| scw | Mon | 4 | Apr | 1994 | 20:40:00 |
| mgb | Wed | 23 | Mar | 1994 | 12:12:21 |
| kmk | Tue | 8 | Feb | 1994 | 09:37:50 |
| mgb | Thu | 12 | May | 1994 | 16:32:51 |
| sjy | Fri | 25 | Mar | 1994 | 19:06:33 |
| sjy | Thu | 24 | Mar | 1994 | 19:51:33 |
| gjfj | Wed | 23 | Mar | 1994 | 10:30:22 |
| daj | Wed | 16 | Mar | 1994 | 00:24:44 |
| jtf | Wed | 9 | Mar | 1994 | 19:18:25 |
| gjfj | Thu | 3 | Mar | 1994 | 12:45:52 |
| mgb | Wed | 2 | Feb | 1994 | 10:29:01 |
| daj | Fri | 6 | May | 1994 | 11:48:44 |
| sjy | Fri | 15 | Apr | 1994 | 12:16:43 |
| gjfj | Wed | 13 | Apr | 1994 | 09:54:12 |
| cbt | Fri | 8 | Apr | 1994 | 16:42:24 |
| jtf | Fri | 1 | Apr | 1994 | 03:18:06 |
| rit | Fri | 1 | Apr | 1994 | 09:56:50 |
| sjy | Mon | 14 | Mar | 1994 | 09:31:53 |
| mda | Wed | 9 | Mar | 1994 | 15:16:23 |
| scw | Tue | 15 | Feb | 1994 | 15:17:10 |

Narrow selection:

□ by users
□ by date
■ by keywords
□ by category

Status: New mail has arrived.

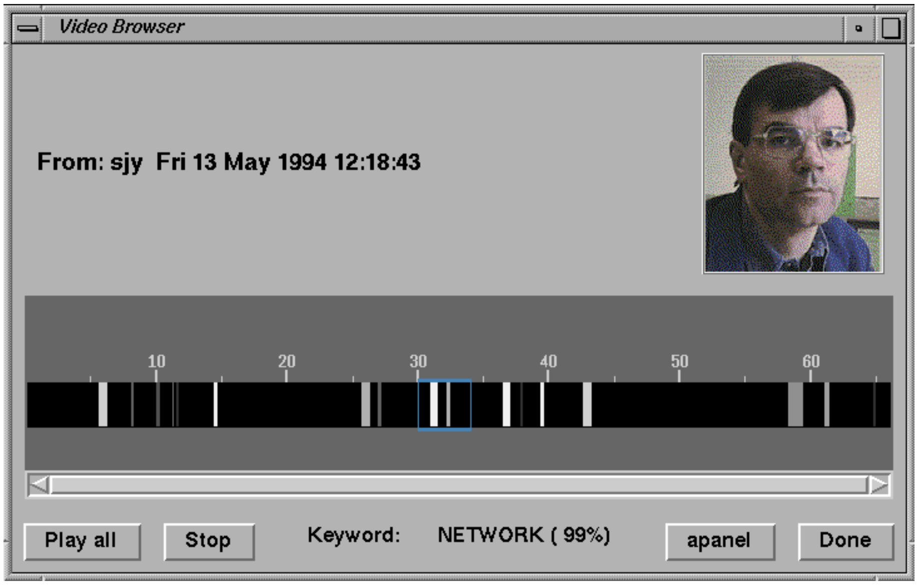
Figure 2: Video Mail User Interface application.

by time. In its simplest form, the keyword search resembles an "audio grep" that returns a list of messages containing a particular keyword.

After the ranked list of messages is displayed, the user must still investigate the listed messages to either find the relevant one(s) or determine that the retrieval was ineffective and that a new search is required. While there are convenient methods for the graphical browsing of text, eg scroll bars, "page-forward" commands, and word-search functions, existing video and audio playback interfaces almost universally adopt the "tape recorder" metaphor. To scan an entire message, it must be auditioned from start to finish to ensure that no parts are missed. Even if there is a "fast forward" button it is generally a hit-or-miss operation to find a desired section in a lengthy message. In contrast, the transcription of a minute-long message is typically a paragraph of text, which may be scanned by eye in a matter of seconds. Clearly there must be more economical ways to access and review audio/video data.

### 3.1 A Video Message Browser

The browser is an attempt to represent a dynamic time-varying process (the audio/video stream) by a static image that can be taken in at a glance. A message is represented as horizontal timeline, and keyword events are displayed graphically along it. Time runs from left to right, and events are represented proportionally to when they occur in the message; for example, events at the beginning appear on the left side of the bar and short-duration events are short. Figure 3 shows a prototype browser. The timeline is the black bar; the scale indicates time in seconds. The brightness of a keyword region is proportional to the spotting score, so that more likely hits appear brighter and stand out (search keywords are shown in a red-black colour continuum while the other keywords are shown in white-black). When pointed at with the mouse, the name and confidence score of a keyword hit are displayed (for example,"NETWORK (99%)" in Figure 1 is the keyword just after the 30-second mark). Portions of the message can be selected for playback by dragging over part of the bar; this lets the user selectively play regions of interest, rather than the entire message. In the figure, a 4-second interval starting 30 seconds into the message is selected for playback.



From: sjy Fri 13 May 1994 12:18:43

Keyword: NETWORK ( 99%)

Figure 3: Prototype mail browser.

### 3.1.1 Video Cues

For this application, we have focussed on the audio stream because that is where nearly all information of practical interest will be found. It is, in general, much more difficult to extract useful information from a video signal[1]. The current state-of-the-art in image retrieval does not extend much beyond relatively crude measures of colour or shape similarity [11, 12]. While efforts in face and gesture recognition are in progress at ORL and elsewhere [13], less sophisticated analyses can still yield information helpful for browsing. For example, a Medusa video analyser module can detect the activity in a video stream; one application uses this activity information to automatically select a preferred video stream from several available in each room. Current plans are to add this activity information to the browser, enabling automatic detection of a camera or scene change. In this case, a "thumbnail" image of the new view would be displayed on the timeline. Also, areas of moderately large activity could be highlighted to indicate that something of potential interest is occurring in the video stream.

## 4 Conclusions and Further Work

Our first retrieval system is admittedly limited in scale, and works mainly on an artificial retrieval task. An important priority is to open the keyword and user sets, allowing a search for arbitrary words spoken by anyone. Though this is a difficult challenge, current work using subword and large-vocabulary recognition is promising [14]. In addition, we are currently recording a large video corpus of BBC news broadcasts, including teletext subtitles, and are investigating more sophisticated methods of retrieving and browsing this information. The first steps presented here suggest that state-of-the-art information retrieval and word spotting techniques can be successfully combined, allowing efficient retrieval and perusal of multimedia information.

## 5 Acknowledgements

Olivetti Research Limited is an industrial partner of the VMR project and we are indebted to them for the use of the Medusa system. This project is supported by the UK DTI Grant IED4/1/5804 and EPSRC Grant GR/H87629.

[1] This is particularly true in the video mail environment, where the vast majority of messages are just "talking head" images from a small pool of users, against static backgrounds.

## References

[1] S. Wray, T. Glauert, and A. Hopper. The Medusa applications environment. In *Proceedings IEEE International Conference on Multimedia Computing and Systems*, pages 265–273, Boston, May 1994. IEEE.

[2] U. Glavitsch and P. Schäuble. A system for retrieving speech documents. In *Proceedings SIGIR '92*, pages 168–176. ACM, 1992.

[3] J. T. Foote, G. J. F. Jones, and S. J. Young. Video Mail Retrieval Using Voice: Report on whole word based keyword spotting. Technical report, Cambridge University Engineering Department, September 1994.

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

[5] G. J. F. Jones, J. T. Foote, K. Sparck Jones, and S. J. Young. VMR report on keyword definition and data collection. Technical Report 335, Cambridge University Computer Laboratory, May 1994.

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

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

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

[9] S. E. Robertson and K. Sparck Jones. Simple, proven approaches to text retrieval. Technical Report 335, Cambridge University Computer Laboratory, Dec 1994.

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

[11] S. W. Smoliar and H. J. Zhang. Content-based video indexing and retrieval. *IEEE Multimedia*, 1(2):62–72, Summer 1994.

[12] R. Barber, C. Faloutsos, M. Flickner, J. Hafner, W. Niblack, and D. Petkovic. Efficient and effective querying by image content. *J. Intelligent Information Sys.*, (3):1–31, 1994.

[13] F. Samaria and S. Young. A HMM-based architecture for face identification. *Image and Vision Computing*, 12(8):537–543, October 1994.

[14] J. T. Foote, G. J. F. Jones, K. Sparck Jones, and S. J. Young. Talker-independent keyword spotting for information retrieval. In *Proceedings of Eurospeech 95.* ESCA, 1995.

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