Spoken document retrieval - a multimedia tool

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

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

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

Abstract

This paper considers some of the issues that arise in multimedia and mode information retrieval, and then describes a current project on spoken document retrieval. The work done so far suggests that users can have retrieval access to information conveyed in a convenient modality that is an increasingly common feature of multimedia systems.

1 Introduction

In this paper I shall first consider the general issues that multimedia or multimode information retrieval present, and then describe some current Cambridge research on retrieval from spoken document files. Multimedia/mode is a global heading for a very large and active area of work, but this effort has concentrated primarily on the presentation of information to the user and on the user's immediate system interface, and core retrieval issues have been largely sidestepped. I shall rather focus on the implications of multimedia/mode for retrieval itself, and on how the lessons learnt from written text indexing and retrieval may apply in the multimedia/mode case. I shall use our current retrieval work, under the Video Mail Retrieval (VMR) Project, to illustrate a first (and modest) application of established text retrieval techniques in the multimedia/mode context. (But I shall only summarise the VMR Project work since it would be improper to repeat, at length, material that has already been published or submitted elsewhere: see (Jones et al. 1995, Sparck Jones et al. 1995).)

2 Some definitions

As e.g. Roth & Hefley (1993) note, there is no real agreement on the use of the terms "media" and "mode" (and hence "multimedia", "multimode" and "multimodal"). This is partly a function of whether things are looked at from the system point of view (tending to bias towards "media") or from the user's point of view (biasing towards "mode"). But it also follows from the complex many-many relations between the two and the difficulty of separating 'what is expressed' from 'how it is expressed'. For present purposes I shall define *medium* as type of information expression, e.g. linguistic, pictorial, and *mode* as manner of information presentation relative to some medium, e.g.written vs spoken language, polychrome vs monochrome picture. The principle is that media are autonomous and not intersubstitutible as means of conveying some underlying information, but that modes are: however this principle is clearly only a relative one and is adopted here simply as an analytical tool. My main purpose in adopting these particular definitions is that they lead us straight into the core issues from a retrieval point of view, by separating internal information storage from external information presentation: these issues remain even if we shift the definitions. Thus if the (logical) document is in some medium, what constraints does this impose first, on how it may be indexed and hence retrieved, and second on how it may, when retrieved, be presented. In particular, if we accept that alternative media are certainly not wholly equivalent as means of embodying and expressing content, even if we treat alternative modes as sufficiently intersubstitutible, and if we also accept that document indexing nowadays is done via requests at search time rather than independently at file time, what do these two factors imply for retrieval? Specifically, what do they imply first for indexing feasibility and effectiveness in relation to the documents themselves, and second for the the user's ability to formulate requests and conduct searches, especially when supplied with a *multi* media or mode interface?

3 Multimedia/mode retrieval issues

We start from the conventional position that users are after document content, while maintaining that content is defined as both information and its expression. The crucial retrieval problem, namely how to state a content need and match for content provision, is then of course the one with which we are familiar from the written document case; but it is more challenging, perhaps very much more challenging, with other media and modes. This can be true even when document and request modes are the same, but will in general be more intractable when not just modes but media differ, and some translation or transformation is required to allow retrieval.

In some cases, e.g. fingerprint matching or perhaps 'mugshot' photofitting, we can take an input request image and search the file for similar images in a very straightforward way. But this is because sufficient overall pattern similarity is required or ensured (say by 'framing'), even if there may be some implicit and weak content characterisation via the form of the similarity definition. In other cases, e.g. searching reconnaissance images, retrieval depends on a prior definition of the content features of interest that uses explicit 'meaning representations' and may therefore support relations between different media or modes. For instance, the language term "missile silo" is translatable into configurations of pixels. In all these pictorial cases retrieval systems, that rely essentially on the notion of likeness between request and document, are highly application-specific. The medium is not intrinsically symbolic, and thus where a strong surface pattern similarity for whole or part images would be far too restrictive to meet actual retrieval needs, request and document cannot be compared without the aid of a bridge. We cannot search an image file directly for the request: find all photographs with sitting small boys in them. At the same time we cannot avoid the need to have a bridge by supposing requests and documents in the same medium and the ability to e.g. point to part of an image and say: find more like this, because it is in general not obvious what 'this' is.

Thus multimedia retrieval systems, when not tightly tied to a specific application, may have to rely on explicit indexing in some indexing language, whatever the medium of the file documents. The indexing language may be uncontrolled natural language, or some controlled language perhaps with a thesaurus, classification scheme etc. In some cases indexing may also use other 'bibliographic-style' keys like dates. These approaches are adopted, for instance, in managing files of television recordings and picture files like those accessed via museum catalogues.

However even where a pictorial medium is already explicitly symbolic, so requests can be couched in the same form as documents and searched directly, as with some graphics languages, this is not necessarily most helpful for the user. It may be more convenient if the user can submit requests in another medium, e.g. natural language, which are then transformed to search the file. But whether or not an independent indexing language is used is then an option, not a necessity as in the previous cases, chosen for the same sorts of reasons as conventional indexing languages for written document retrieval.

We have an analogous situation when dealing with modes rather than media, even if mode differences are smaller than media ones. Even where content is symbolically represented, translation may be required, just as English requests might be translated into German or French for searching multilingual written files. Alternatively, translation may be convenient though not neccesary, for instance it may be more convenient, in a 'hands-occupied' situation, to accept spoken requests for searching written files. However as these examples suggest, mode crossing may be inhibited by the present state of technology. Thus it is not difficult, with current speech recognition, to accept and transcribe spoken requests if these are carefully uttered. But spoken language files, even though already in principle holding symbolically-encoded information, are not generally or readily accessed using current speech recognition technology, and retrieval from spoken document files is thus a major challenge.

This is an important problem that needs to be addressed. First because, however important other media are, natural language is itself a vital means of communicating and seeking information; but second because, however important written language is and will remain, spoken language files are becoming more voluminous and more significant. Moreover even if explicit indexing was deemed desirable from a retrieval point of view, it is not an economical or even practicable proposition given the large volumes and low uses of much spoken material and the sheer time required to listen to the documents in order to index them.

4 The VMR Project

The Cambridge VMR Project is attacking spoken document retrieval by combining state-of-the-art speech recognition technology with state-of-the-art natural language indexing techniques; the project is, moreover seeking to profit from the fact that these derivative, request-driven indexing techniques are not just the only practicable ones but have independently-proven merits. The project has, however, to demonstrate that these strategies, developed for the written case, are both viable for and suited to the spoken case. In relation to the general issues of multimedia and mode retrieval discussed earlier, spoken language retrieval, however technologically difficult, is less intractable from the point of content characterisation and access that e.g. image retrieval. But it is not trivial as a form of retrieval, and established speech retrieval strategies would be very useful.

The VMR Project, moreover, has multimedia and mode interest not just from the speech retrieval point of view in itself. The project is motivated by Olivetti Research Ltd (ORL)'s Pandora and Medusa systems. The original Pandora, and successor Medusa system now being built, are pre-eminently multimedia systems in terms both of the types of material they can handle and of the ways in which the user can interact with computing and communication facilities through a sophisticated multimedia and mode interface. The VMR Project is thus as much an initial exploration of the role of spoken language retrieval within a user's multimedia world, supported by a multipurpose workstation, as of speech retrieval for its own sake.

ORL's Pandora and Medusa systems offer both interactive video and video mail, and video archives can grow very rapidly, prompting the same interest in retrieval both from an individual user's files and from the community's files as email has done. However while it would in principle be nice, for full multimedia retrieval functionality, to be able to search the visual as well as audio files for the video material, this is currently neither feasible for the kinds of reason mentioned earlier nor probably very useful in this particular case. Thus the visual images consist primarily of speaker pictures, from a known community; and while such view of individuals can provide additional information, the main information content in the current files is in the spoken record. We are therefore concentrating on speech retrieval and will present automatically correlated video images only to accompany spoken messages output by the retrieval system. We are, further, working only with the video mail material, since the more informal video dialogue is well beyond the capabilities of present speech recognition techniques. The video mail messages are generally somewhat more formal, though they are still more natural than the read or dictated material used for e.g. the ARPA non domain-specific speech recognition evaluations (Kubala et al. 1994).

5 Speech retrieval problems

Spoken document retrieval, though natural language is the medium, presents distinctive, non-trivial problems for retrieval. The first is that the sheer bulk of the acoustic record, even when compressed, means that very efficient internal processing is required. The second is that the spoken modality leads to documents with different discourse and linguistic properties from written ones. The third is that speech is a slow modality, so efficient presentation of retrieved material is much more of an issue than with written documents. The fourth is that automatic speech recognition technology for natural (as opposed to dictated or read) speech is far from perfect, so no speech retrieval system can have the grounding that written document ones do, in reliable and complete incidence data for the words in the file.

More specifically, the retrieval problems of the VMR application can be summarised as follows. They are partly those of speech retrieval in general, partly those of this application task. Both have implications for the attempt to use established (statistical) text retrieval methods.

The first class of problems concerns the indexing foundation, namely individual terms. Word identifcation is not certain, and pronunciation differences for word forms can make it less easy to pick up common stems than in the written case. Speech recognition assigns an acoustic score, or *a-score*, to words, so e.g. thresholding has to be judged for its effect on *hits*, *misses* and *false alarms* in word identification. On the other hand while actual a-scores, because they are probabilistic, can in principle be combined with other probablistic weighting factors of proven utility (Robertson & Sparck Jones 1994), it is not obvious how to do this in practice. (One particular point it that the within-document occurrences of a term may have different a-scores.)

The second class of problems concerns the documents, and the distinctive properties of spoken as opposed to written discourse, e.g. greater informality. In the VMR case there are also the specific properties of email-like messages to be considered where, for example, messages may be brief and/or elements in chains, and contain virtually no content words. Then when the uncertainties of word recognition are combined with the sheer bulk of the files, it becomes impossible (or certainly impracticable) to gather the distributional data characterising the entire file on which statistically-based retrieval methods depend. There can, for example, be no question of having an inverted file for the entire word vocabulary along with text-position indicators. This has implications for the underpinning mechanics of retrieval and imposes constraints on strategies like query expansion.

There are also both project-independent and project-dependent constraints on the user interface. Listening to speech is much slower than scanning text, so output strategies have to be designed e.g. to provide effective quotations from retrieved documents: current speech recognition technology is not good enough for transcription even of selected output. At the same time, given the multimedia/mode interface for the Medusa system, information about the spoken messages given by their email-like (written) headers and their associated video images has to be effectively managed and used in conjunction with the spoken message *bodies* themselves.

6 VMR Project strategy

There has been very little work on speech retrieval. There has been much more on word spotting, i.e. identifying occurrences of specific words in a speech stream. Of course speech retrieval has to look for term occurrences in the file, but conventional word spotting as practised by e.g. by the intelligence community is not necessarily designed to meet normal document retrieval criteria, and there has in any case been little publicly reported performance evaluation. Reported work on word spotting for e.g. categorisation (Rose 1991, McDonough et al. 1994), on the other hand, is usually intended to meet document classification needs that are much coarser than ordinary retrieval ones. Work on speech retrieval in the present sense has been done by Glavitsch & Schäuble (1992), but so far via simulation, and the research by Kupiec, Kimber & Balasubramanian (1994) has been on spoken access to written text files. At the same time, though performance for continuous speech recognition has improved significantly over the last decade, the ARPA tasks use read speech or are within limited domains, so the results obtained, especially for former as more relevant than the latter, are still only suggestive for and not directly applicable to the VMR case. However the speech recognition engine, HTK, used as the basis of the VMR work, has demonstrated good performance in the ARPA read task and is thus a very suitable vehicle for the project work (Young, Woodland & Byrne 1993).

Our project strategy is therefore an incremental one with three phases of work aimed at successively more powerful (or less limited) systems. The first uses a fixed keyword vocabulary, a known user community, and simple interface; the second opens up the speaker set and the third the search vocabulary, accompanied by interface developments and the use of other search keys from message headers as well as words from message bodies.

7 Test data

One of the problems in testing speech recognition performance is having the data to do this: this covers not only suitable input material, but 'answer data' in the form of text transcriptions. We have not been able to use existing ORL material: Pandora recording was with too low a quality microphone, and while this is not a problem with Medusa, there is not yet any available Medusa data. We have therefore engaged in a test data construction exercise designed to capture as far as possible both the content distribution properties of real document files and the kind of speaker characteristics expected: thus the test document set cannot consist of a wholly arbitrary set of items but must have the sort of similarities and dissimilarities that search strategies have to deal with. At the same time since transcription, necessary for calibration, is expensive, the data could not be large and tests with it are only initial feasibility ones.

VMR Database 1 is described in detail in Jones et al. (1994). Briefly, we constructed it using a set of 35 *keywords* and 10 broad subject *categories*. The keywords and categories were chosen both to be sensible for available speakers and to ensure overlaps in the usage both of keywords and other words across categories. We obtained 300 messages, 30 per category, via *prompts* that sketched scenarios designed to stimulate messages falling within the category and using the associated keywords in a natural way. To ensure both message topic and speaker variation we had 5 prompts per category and sought messages from 6 speakers for each. We had 15 speakers altogether, with the individual speakers assigned in semi-random fashion across 4 categories each: they included men and women with varying backgrounds and interests. Both to obtain initial messages with good acoustic properties and to test the effects (relevant to the eventual Medusa environment) of microphone position, we recorded in a soundproof room with head and desk microphones in parallel. The messages in the set vary in speech and language formality, and in length, as well as in topic. They were transcribed using standard protocols, so the the nature of the messages in our speech file can be illustrated by a transcribed example as follows:

> I understand you're having some problems with the synchronisation of acoustic data [pause] on the indigo [loud_breath] with [ah] keyboard control. [loud_breath] [um] Sounds like a software problem to me. [um] [loud_breath] My initial reaction would be to suspect a direct memory access blocking some other key process [loud_breath] which is required to [er] keep the two in synch. [loud_breath] [um] What I suggest you do is have a word with [er] whoever wrote the [loud_breath] [um] the driver software for the [er] [pause] those two operations, both the [ah] acoustic input and the keyboard control, [loud_breath] [ah] and check to see that [um] [pause] one process is not being blocked by another.

There is only one keyword in this message, indigo, but also words that might be confused with keywords, like keyboard with keyword. (The punctuation is inserted by the transcriber to aid reading, and is not automatic.)

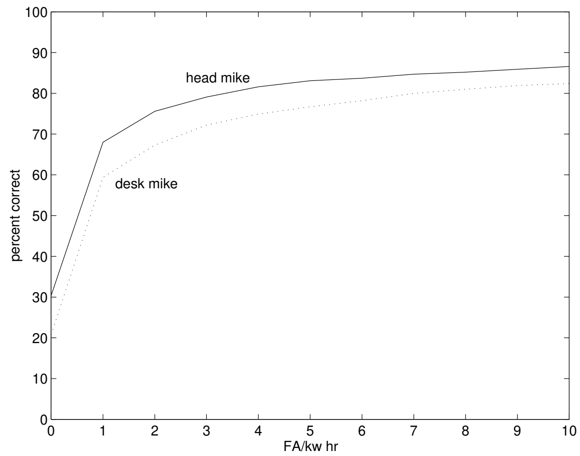

Figure 1: VMR Task Keyword Spotting ROC

8 Retrieval tests

We have already conducted two sets of retrieval tests with this data and are midway through a third.

While we intend to use, as far as is possible in the speech case, established statistically-based techniques, we have to tackle the two classes of problem mentioned earlier: first, how to manage searching given the lack of prior term incidence data; and second, how to combine information about terms derived from acoustic processing with other information, especially for weighting purposes. However in our first stage work we are using a fixed keyword vocabulary and have therefore been able to create inverted files for this (which are accessed by the written not spoken form of the words); we have moreover been able to take advantage of the flexibility of acoustic matching and subsume most of the occurring variant forms of a keyword under its paradigm form, in a kind of pseudo stemming. At the same time, to keep our first tests simple, we have not made any attempt to use a-scores directly in retrieval, but have applied a threshold defining term presence. We have in fact applied a range of a-score thresholds, with lower ones permitting more word occurrences, in order to study the effects of acoustic matching on retrieval. We thus have a set of inverted files rather than just one.

8.1 Word spotting

The simplest form of retrieval is that defined by *keyword spotting* with one keyword per 'request'. Output for the set of keywords can then be used to establish speech recognition performance, independent of request formulation factors: this follows conventional approaches based on hits and false alarms and defines an operating characteristic curve for the speech processing system under changes in the a-score threshold. The transcription data is used for measurement. As well as giving basic word matching performance, these tests supplied information about the comparative performance of the two microphones (and indeed were also used to generate the inverted files). Figure 1 gives the results for the 35 keywords as the percentage of hits when the a-score threshold is varied to allow from 1 to 10 false alarms per keyword per hour. This shows both a good level of speech recognition and a performance level for the desk microphone (more acceptable to likely users) not far from that of the head microphone.

Weight Scheme		Head				Desk			
		cl		idf		cl		idf	
		th_{mco}	th_{map}	th_{mco}	th_{map}	th_{mco}	th_{map}	th_{mco}	th_{map}
Precision	5 docs	0.248	0.232	0.288	0.256	0.252	0.244	0.264	0.260
	10 docs	0.212	0.192	0.218	0.222	0.176	0.182	0.226	0.214
	15 docs	0.164	0.169	0.185	0.195	0.164	0.167	0.197	0.191
	20 docs	0.143	0.156	0.160	0.171	0.141	0.142	0.172	0.166
Av Precision		0.256	0.259	0.293	0.295	0.232	0.241	0.283	0.283

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

	Text	Phonetic	Head	Desk
Relative to Text	100%	95.5%	88.8%	85.2%
Relative to Phonetic	—	100%	93.2%	89.4%

Table 2: Relative average precision retrieval performance with idf term weighting.

8.2 Message searching

Our second set of tests were much less narrow and marked our first step towards true retrieval experiments. They could not, however, be genuine as we have no real users or requests for Database 1. We therefore derived requests and 'relevance assessments' as follows. We treated each prompt as a request, and extracted all the keywords in these to form, after stemming, the search query. For example, one prompt and its query were:

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

As the foregoing implies, we are working with written queries, but this is sensible in the first instance and does not invalidate our speech retrieval goal. We then took the messages generated for each prompt as their relevance set: they were not always very closely related to the 'query' but were sufficiently so to support experiments with message retrieval, particularly since other messages for the same category but different prompts were deemed non-relevant and had thus to be discriminated against in searching, even though they shared keywords with the relevant documents.

In the set of runs we have carried out so far we have treated plain term matching by coordination level (*cl*) as a baseline, and taken the first step towards weighted retrieval using inverse document frequency weights (*idf*). Since our message set is small we have used precision at document output cutoff up to rank 20, and also average precision (both as defined for the TREC tests (Harman 1994)) to measure performance. Again, in order to study recording effects, we computed results for both head and desk microphones.

In order to *calibrate* performance, as is necessary for the speech retrieval case, we proceeded much as in the word spotting tests and compared spoken message output with that obtained by doing normal written text retrieval on the transcribed messages: this defines the comparative upper bound for the speech case. However since some orthographically distinct words are pronounced in the same way, phonetically transcribed text supplies a fairer upper bound. Finally, for these first tests, we assumed we had chosen the best possible keyword occurrence threshold: that is we used the best *a posteriori* thresholds for output cutoff 5 and for average precision, respectively th_{mco} and the_{map}, as if they had been set *a priori*. This is not strictly legitimate, but is a reasonable simulation of threshold setting based on extensive operational experience.

The results we obtained are given in Tables 1 and 2. These again show desk microphone performance not far behind that for the head microphone; performance for *phonetic* text as slightly inferior to that for ordinary *text*; and the expected gains with *idf* weighted as opposed to *cl* queries. The absolute performance figures are of course not indicative of real levels of performance with genuine requests and relevance assessments, or with a non-trivial size of document file. They do, however, show speech recognition performance able to support adequate levels of retrieval performance compared with the written text case.

9 Current and future work

The experiments currently in hand involve assembling real requests and accompanying relevance assessments from users. Our next steps will be to remove the restriction to a known speaker set: the speech recognition system was tuned for the first tests using specially recorded separate training data for the message speakers; but there is enough general-purpose training data available now for testing without speaker tuning. Our second step will be to replace Database 1 with a larger, real document set. The third step will be to develop the user interface. By the end of Stage 1 we had a comparatively simple but attractive interface able to provide a graphic display showing the location (and hence any concentration) of query terms in a message, so it was possible to home into that part of the recording (Brown et al. 1994). However we need to do more work on the presentation of search output to the user and on appropriate supports for user follow-up searches.

The spoken document retrieval work just described addresses only a relatively tractable manifestation of the multimedia and mode retrieval problems considered earlier. It is concerned with a case where the information communicated is already symbolic but is somewhat noisily transmitted, and where mode shifting is relatively unproblematic. The problems to be overcome even in this comparatively favourable case are nevertheless far from trivial. At the same time, if spoken document retrieval is feasible, and the work done so far on the VMR Project suggests it may be, it can make major contributions, from the user's point of view, to efficient and effective multimedia and mode systems.

10 Acknowledgements

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

References

Brown, M. G., Foote, J. T., Jones, G.J. F., Sparck Jones, K. & Young, S. J. (1994), Video mail retrieval by voice: an overview of the Cambridge/Olivetti retrieval system, *in* 'Proceedings of the ACM Multimedia Database Workshop.

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

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

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

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

Kubala, HF. et al. (1994), The hub and spoke paradigm for CSR evaluation, *in* 'Proceedings of HLT 94', (ARPA), San Francisco: Morgan Kaufmann, pp. 37–42.

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

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

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

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

Roth, S.F. & Hefley, W.E. (1993), Intelligent multimedia presentation systems: research and principles. *in* M.T. Maybury, ed., *Intelligent Multimedia Interfaces*, Menlo Park: AAAI Press, pp. 13–58.

Sparck Jones, K., Foote, J. T., Jones, G.J. F. & Young, S. J. (1995), Retrieving spoken documents: Stage 1 of the VMR Project, submitted to ACM-SIGIR Conference, 1995.

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