ACOUSTIC INDEXING FOR SPOKEN DOCUMENT RETRIEVAL

*Gareth J. F. Jones**

Department of Computer Science, University of Exeter, Exeter EX4 4PT, U.K.
Tel: +44 1392 264061 Fax: +44 1392 264067 email: gareth@dcs.ex.ac.uk

ABSTRACT

The retrieval of spoken documents presents new problems in information retrieval and browsing. This paper reviews recent experimental research in the automated retrieval of spoken documents and outlines an operational prototype which demonstrates that effective vocabulary-, speaker- and domain-independent spoken document retrieval is now possible.

1. INTRODUCTION

Large archives of digitally stored audio and video data are becoming increasingly common. Locating information within such archives poses new problems in information retrieval and browsing. For example, a fundamental question is how to index the contents of multimedia documents. The development of efficient techniques for the automated retrieval of multimedia documents is particularly important since manual browsing of audio and video material is inefficient and very time consuming.

The development of automated multimedia retrieval systems requires the integration of various technologies including multimedia hardware and software, information retrieval and content indexing. Many systems have been developed for the capture, processing and real-time playback of multimedia data. This paper concentrates on advances and ongoing research in retrieval technology of spoken documents and outlines an operational prototype spoken document retrieval (SDR) system.

2. INFORMATION RETRIEVAL

Information retrieval (IR) techniques are used to satisfy a user's information need by retrieving documents from unstructured collections. In retrieval the user typically enters a search request; in response to this the IR system returns a set of potentially relevant documents taken from the collection. Depending on the IR system used, the search request may either be in natural language or use some more formal Boolean structure. In addition, the returned documents may often be usefully ranked, based on a matching score between the query and the document, with the documents judged most likely to be relevant at the top of the list. For text documents the user's information need is satisfied simply by reading the returned documents until they find the information they are looking for.

3. INDEXING SPOKEN DOCUMENTS

Text retrieval represents the most straightforward IR environment since the contents of the documents are available for analysis. For SDR the contents must first be indexed using speech recognition techniques. Unfortunately automatic speech recognition is not completely reliable and all current speech recognition systems have technical limitations. SDR research has investigated various approaches to indexing including the following:

Word Spotting: A *word spotting* (WS) system uses speech recognition to attempt to identify occurrences of predefined keywords in spoken documents [1]. Recognition using this technique is generally quite reliable, but significant drawbacks are that the keyword vocabulary is typically limited to less then 100 words, and that all documents must be recognised again if a new keyword is added. While WS may be useful for classification of documents into fixed categories, it will usually not be effective in ad hoc retrieval where users are free to form new search requests.

Phone Lattice Spotting: An alternative WS technique is *phone lattice spotting* (PLS) [2]. In PLS a sub-word lattice in the form of an acyclic-graph is constructed in the main speech recognition pass. When a search word is entered the sub-word lattice is scanned for sub-word unit sequences corresponding to the search word. PLS does not have the vocabulary constraint of standard WS, but recognition is generally less accurate.

Large Vocabulary Recognition: *Large vocabulary recognition* (LVR) attempts to completely transcribe the contents of a spoken document. However despite recent rapid advances in LVR technology, these systems still make mistakes and are unable to recognise words outside their vocabulary. Thus even with a vocabulary of 60,000 words many useful search words, particularly proper nouns, will be outside the vocabulary.

Subword (n-gram) Indexing: A vocabulary independent technique for spoken document indexing is to use n-gram sequences of subword units [3]. For retrieval n-grams are extracted from the search words entered by the user and matched against those found in the documents. n-gram indexing is particularly attractive for generative languages such as German where new compound words, which will be outside the vocabulary of an LVR system, occur very commonly.

4. EXPERIMENTAL STUDIES

A number of studies have investigated spoken document retrieval. The first research used WS with a topic-classification technique [1]. A system similar to this has recently been developed for Japanese news classification [4]. The first project to propose the use of classical information retrieval techniques in spoken document retrieval was [3], where subword indexing units were used to retrieve Swiss German news stories. The use of PLS indexing was proposed and investigated in [2]. LVR for spoken document retrieval is being investigated in the Informedia project [5]. A comparative study of all these indexing techniques has been carried out within the Video Mail Retrieval project at Cambridge University and ORL, Cambridge [6], in which

* Currently: Visiting Fellow, Research and Development Center, Toshiba Corporation, 1 Komukai Toshiba-cho, Saiwai-ku, Kawasaki 210, Japan

I participated. The VMR project demonstrated that effective speaker-independent domain-independent spoken document retrieval is possible. Our experimental results on a small archive of 300 voice mail messages showed SDR performance using PLS or LVR of around 80% of that achieved using manual text indexing of the messages, and also that combining indexing techniques can result in improvement in retrieval performance to around 90% of that for text.

4.1. TREC Spoken Document Retrieval Track

Interest in SDR has increased greatly in recent years and in 1997 for the first time an SDR track was included in the annual US NIST TREC (Text REtrieval Conference) [7]. Although still a comparatively small task with less than 2000 documents in the retrieval collection, this represented the first time different researchers have had access to a common SDR collection. For each document the TREC organisers provided manual text transcriptions and LVR transcriptions generated by IBM. Work as part of the City University effort [8], applying the IR techniques used in the VMR project to the IBM transcriptions, showed that these scaled up successfully and performed well in comparison with results obtained by other teams using either the IBM transcriptions or their own speech recognition output.

5. THE VIDEO MAIL RETRIEVAL SYSTEM

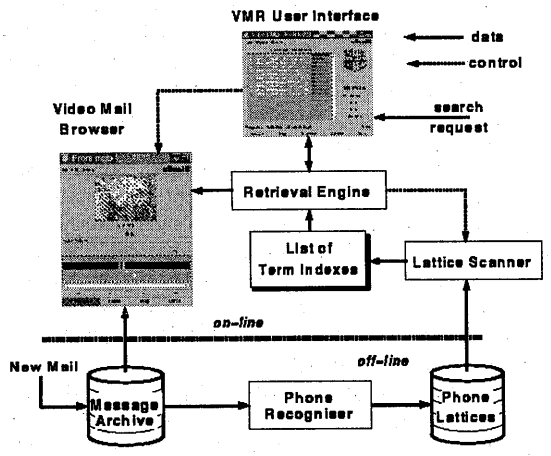

Figure 1: Block diagram of video mail retrieval system

In the VMR project we developed an effective prototype SDR application, described in detail in [9]. Figure 1 shows a block diagram of the VMR system. The prototype system uses only PLS indexing. When new messages arrive speech recognition is performed to generate a phone lattice. At retrieval time when a user enters a query the lattice is scanned for occurrences of each query word. After PLS is complete a query-message matching score is computed for each message, and a ranked list of messages is generated and displayed to the user in the interface shown in Figure 2.

Since browsing by listening to complete spoken messages from the beginning is very time consuming we developed a graphical browser application, shown in Figure 3. A message is represented as horizontal timeline, and recognised search terms are displayed graphically along it. The user can start message playback at any time simply by clicking at the desired point in the time bar; this lets the user selectively play regions of interest, rather than the entire message.

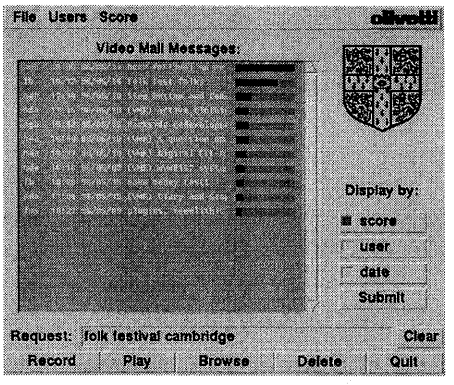

Figure 2: Video mail retrieval GUI

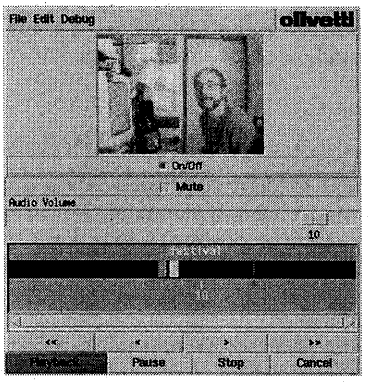

Figure 3: The video mail browser

6. REFERENCES

[1] R.C.Rose. *Techniques for Information Retrieval from Spoken Messages* Lincoln Laboratory Journal, 4(1), pp45-60, 1991

[2] D.A.James. *The Application of Classical Information Retrieval Techniques to Spoken Documents* PhD thesis, Cambridge University, 1995

[3] U.Glavitsch and P.Schäuble. *A System for Retrieving Spoken Documents* Proceedings of the 15th ACM SIGIR Conference, pp168-176, Copenhagen, 1992

[4] Y.Ariki and M.Sakurai *Classification of News Speech Articles by Keyword Spotting* Proceedings of the Acoustical Society of Japan and Acoustical Society of America Third Joint Meeting, pp1015-1020, 1996

[5] A.G.Hauptmann and M.J.Witbrock *Informedia: News-on-Demand Multimedia Information Acquisition and Retrieval*, in Intelligent Multimedia Information Retrieval (ed. M.T.Maybury), pp215-240, The AAAI Press/The MIT Press, 1997

[6] G.J.F.Jones, J.T.Foote, K. Sparck Jones and S.J.Young. *Retrieving Spoken Documents by Combining Multiple Index Sources* Proceedings of the 19th ACM SIGIR Conference, pp30-38, Zurich, 1996

[7] D.K.Harman and E.M.Voorhees (editors) *The Sixth Text REtrieval Conference (TREC-6)* Gaithersburg, MD, NIST, 1998

[8] S.Walker, S.E.Robertson, M.Boughanem, G.J.F. Jones and K.Sparck Jones. *Okapi at TREC-6: automatic ad hoc, VLC, routing, filtering and QSDR* in [7]

[9] M.G.Brown, J.T.Foote, G.J.F.Jones, K.Sparck Jones and S.J.Young *Open-Vocabulary Speech Indexing for Voice and Video Mail Retrieval* Proceedings of ACM Multimedia, pp307-316, Boston, 1996