AT&TV: Broadcast Television and Radio Retrieval

Timothy J. Mills, David Pye, Nicholas J. Hollinghurst & Kenneth R. Wood

AT&T Laboratories Cambridge,
24a Trumpington Street, Cambridge CB2 1QA, England

dart@uk.research.att.com

ABSTRACT

This paper reports recent work at AT&T Laboratories Cambridge to develop retrieval systems for broadcast television and radio programmes. Unlike some other systems, it does not rely on manual classification or annotation of the broadcast material; it is indexed automatically from the air. While many digital video library projects focus solely on broadcast news, we have broadened our efforts to produce a continually updated index of all UK terrestrial TV output.

We have investigated the use of both video and audio stream analysis, together with closed-caption and speech recognition derived text transcripts, to aid retrieval of whole programmes and of segments within them. We report on the design of two prototypes. The first employs speech recognition to generate transcripts for indexing of radio and television news broadcasts. The second system entitled AT&TV uses closed-caption transcripts where available, to index a much wider range of programmes. A centralised archive is maintained of the past seven days television across four terrestrial channels, thus removing a viewer's need to select programmes prior to broadcast. The system delivers video-on-demand to clients with broadband access and is in regular use in a lab-wide deployment with around 50 users.

1. INTRODUCTION

The DART (Digital Asset Retrieval Technology) project [URL 1] at AT&T Laboratories Cambridge is concerned with the management of digital media including text and hypertext documents, images, audio and video recordings. DART aims to provide the means to index, annotate, navigate and retrieve from diverse collections of these assets. In part, the project follows on from our successful collaboration with Cambridge University in the Video Mail Retrieval project (Jones *et al*, 1996). Whereas VMR relied solely on speech recognition of the soundtrack, the DART project has advanced these techniques and combined them with content-based image retrieval, acoustic searching and video parsing techniques.

In developing the core DART technologies, two prototypes for indexing broadcast television and radio material have been constructed. The first of these is a news retrieval system, which uses automatic speech recognition (ASR) to generate transcripts for input to a text indexer. The system can handle both radio and television news broadcasts. The second system goes beyond the limited domain of news broadcasts to encompass an entire week of British terrestrial television, updated throughout each day. It makes use of closed-caption transmissions, where available, instead of ASR.

The paper is structured as follows. Section 2 discusses related work. Sections 3 and 4 describe the data capture and analysis processes and the resulting representation of programme content common to both applications. Sections 5 and 6 describe particular techniques used or investigated to facilitate news retrieval and general TV retrieval respectively. We end with a discussion and conclusions in section 7.

2. RELATED WORK

Among the first projects to investigate use of speech recognition for information retrieval was the Cambridge Video Mail Retrieval project (Jones *et al*, 1996). Initially using keyword-spotting and later large vocabulary speech recognition, VMR showed that even with imperfect speech recognition, information retrieval was still highly effective. The same group produced a news retrieval system (Brown *et al*, 1995) based on closed-captions (rather than speech recognition) to experiment with multimedia archives. To capture and record TV broadcasts it used Medusa (Way *et al*, 1994), an

experimental distributed multimedia system with a specialised high-performance hardware infrastructure. We have expanded upon, scaled-up and made robust this experimental work to produce a news retrieval system using ASR, and a multi-channel TV archive that are in daily use. The feasibility of broadcast media retrieval has been well established — we wished to demonstrate the usefulness and scalability of the technology.

Many academic and industrial research groups have developed digital video libraries. The Infomedia project at Carnegie Mellon University has published a large amount of material on various techniques for searching and browsing broadcast media, mainly focusing on news programmes [URL 3]. The techniques include use of speech recognition and video image analysis. The VISION Digital Video Library System at the University of Kansas (Gauch *et al)* features a pipelined digital video processing architecture which is capable of digitizing, processing, indexing and compressing video in real time on an inexpensive general purpose computer. AT&T Laboratories in New Jersey have also done a substantial amount of work on techniques for searching and browsing video (Gibbon *et al,* 1999) and have recently begun building systems aimed at broadcast quality MPEG-2 video. Much of this work is also applicable to general TV material, on which we have concentrated our efforts.

The speech recognition community has also made a significant contribution to work in this area. In particular, NIST's Broadcast News Evaluation has over recent years, helped participants such as Cambridge University Engineering Department (Woodland *et al,* 1998) and LIMSI to make significant progress in audio indexing performance. The SpeechBot team at Compaq has produced an on-line demonstration of searching a large archive of indexed radio programmes [URL 6].

Digital TV recording is now a commercial reality. TIVO [URL 7] and ReplayTV [URL 5] both offer consumer digital TV recording using MPEG-2 compression and hard-disc technology. However, the storage space is limited and so it is necessary to select in advance which programmes are to be recorded. Some of the recordings are made automatically, using metadata supplied by the service provider and matched against a profile of the users' interests. Our approach uses a much larger amount of centralised storage, relying on broadband networks to distribute the video on demand. This enables us to store a whole week's television output and provide content-based programme filtering, browsing and searching of the entire archive.

There is now some commercial activity in on-line video searching. MediaSite [URL 4] provides a video indexing service using image, audio and closed-caption content. It is aimed at content providers making available stock footage, and allows human interaction in the indexing process to select video segments and apply annotations. We have opted for a fully automatic system that can deliver either whole programmes or video clips.

FasTV.com [URL 2] is a web-based service that aggregates a selection of current and archived video as indexed segments averaging 90 seconds in length. Its content appears to be largely news-based and the service is not intended for whole-programme viewing.

3. SYSTEM ARCHITECTURE

The five British terrestrial television channels are accompanied by teletext, a data information service transmitted together with normal television signals. This includes a programme schedule giving titles, times and information about the day's broadcasts and a programme delivery control (PDC) signal used by suitably equipped video recorders to handle unexpected delays in transmission. For the majority of television programmes, a closed-caption service is provided for the benefit of the hard of hearing.

Figure 1 shows the architecture of the digital television recording system. The system comprises four main processes: capture, parse, index and search and their associated data sources and storage. TV broadcasts are captured as five 1.4Mbit/s MPEG-1 streams (600Mbyte/hour) using a Telemedia Systems [URL 7] networked MPEG-1 encoder for each channel. MPEG-1 was chosen in the applications for its relative compactness and accessibility from both Unix and Windows-based

clients. The system uses one teletext decoder for each channel to receive the closed captions, plus a further decoder to receive the programme schedules. The output is sent via an ATM network to a Sun Enterprise 4000 that places the data on a half-Terabyte RAID filesystem. Recording proceeds according to the programme schedule, with each programme being stored in a separate file.

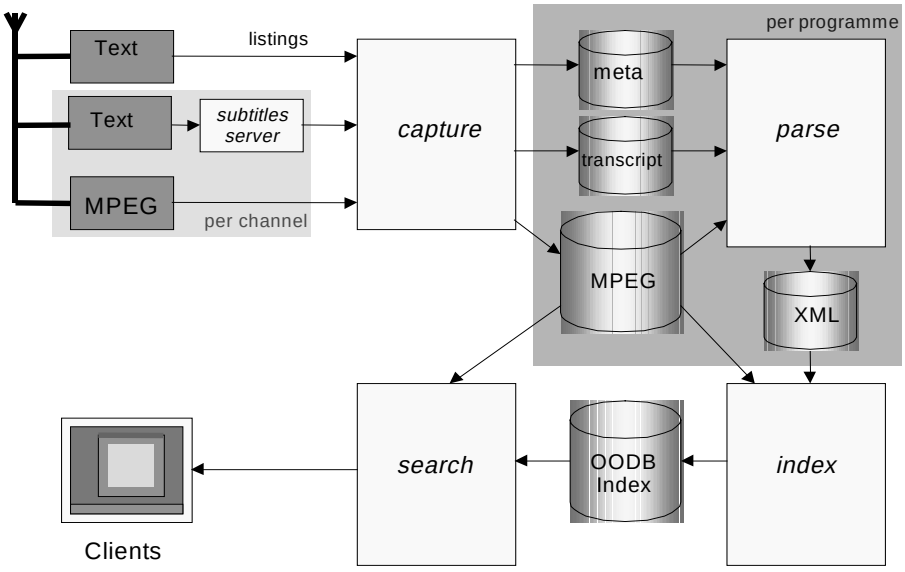

Figure 1. System Components

As each recording completes, a process is initiated to parse the MPEG audio and video streams along with the closed captions. Audio and video are analysed independently and the final segmentation and content information is computed from the results. This process generates the XML representation of programme content described in Section 4. Table 1 shows typical processing times for the News system per hour of material. In the general TV system, the first three stages are identical and the ASR stage is omitted.

	CPU time (hours)	Real time (hours)
Capture MPEG	0 (hardware)	1:00
Audio decode/analyse	0:24	1:00
Video decode/analyse	1:00	
ASR	7:00	1.18

Table 1. Typical processing time for one hour of TV on a 6 CPU 166MHz Sun Enterprise 4000

The XML document is then presented to the information retrieval system for indexing. Our IR system is a modular object-oriented database system (OODB) with specific support for IR tasks (Mills *et al,* 1997) such as an inference network model for retrieval (Turtle, 1991). The OODB contains objects representing the content of each programme, together with inverted text indices for rapid searching of transcripts. A web server was constructed to query the OODB and generate HTML pages for display on remote clients. It can also extract and transmit part or all of a relevant MPEG file to a client on demand.

4. CONTENT REPRESENTATION

We represent the content of radio and television programmes in a structured, hierarchical way. This structure is manifest both in our intermediate XML representation of content and in the referential structure of our object-oriented database. It also corresponds closely to the structure and linkage of the web pages generated by our search engine.

The top-level element is the programme, which carries metadata such as listings information and the duration of the programme. Each programme is subdivided into segments. These segments form the basic unit of information retrieval and are the objects retrieved using the inverted text indices. A typical programme length is around half an hour, which is rather too long for rapid browsing. Many programmes are magazine-style; they contain a number of short articles on different topics. Clearly, the retrieval system should aim to pinpoint the segments of such programmes rather than retrieving the programme as a whole. If segments are too long, they might span unrelated items and present irrelevant material to the user. On the other hand, very short segments might not contain enough relevant words for retrieval and could result in an incomplete, fragmented presentation.

The approach taken to segment the programme determines coincidences between acoustic boundaries and video shot-breaks (Pye *et al,* 1998). The acoustic boundaries are produced by an algorithm originally developed to reduce soundtracks of arbitrary size into manageable portions for speech recognition. The algorithm splits the audio stream at points where the acoustic characteristics change markedly. These points typically signify changes in speaker, microphone or acoustic channel conditions, or the starting or finishing of music. Since this part of the audio processing is fast, it is used regardless of whether or not the soundtrack is to be subjected to speech recognition. Each segment's acoustic parameters are used to determine whether it is predominantly music. If so, the segment is omitted from the speech recognition process.

The video stream is also parsed to determine the logical structure of the video stream by detecting cuts, fades, dissolves and camera motion. In addition, a representative keyframes is assigned to each segment. A simple block-histogram method is used to detect video change and a series of temporal models then detect characteristic cut and dissolve events. Camera motion is recovered using robust statistical analysis of MPEG-1 motion vectors, although this is not currently used for audio/video segmentation.

For TV programmes, the resulting audio and video breaks are combined using an iterative algorithm which selects acoustic boundaries to form suitably sized segments, favouring audio boundaries which are close to video breaks. Subjectively, this provides a better segmentation of items in news and magazine shows than segmentation based upon audio or video breaks alone. For radio programmes, video breaks are of course not available and therefore audio breaks alone are used. Given a preliminary audio/video segmentation, an optional post-processing stage uses information retrieval techniques to merge adjacent segments of similar lexical content. A set of keyframes is then selected from the video content to represent each segment visually.

Each segment may also contain one or more speech units. These are segments determined by the audio analyser to contain speech from a single speaker and environment. Each speech unit is stored containing a list of words and its acoustic parameters to enable retrieval of similar sounding clips.

Each word is tagged with a start and end time determined either during speech recognition or from its closed-caption timestamp.

The tag structure shown in Figure 2 can thus describe the content of both television and radio programmes. The information retrieval system need therefore only be designed to process XML documents with this structure, without concern for the input medium or means of transcription.

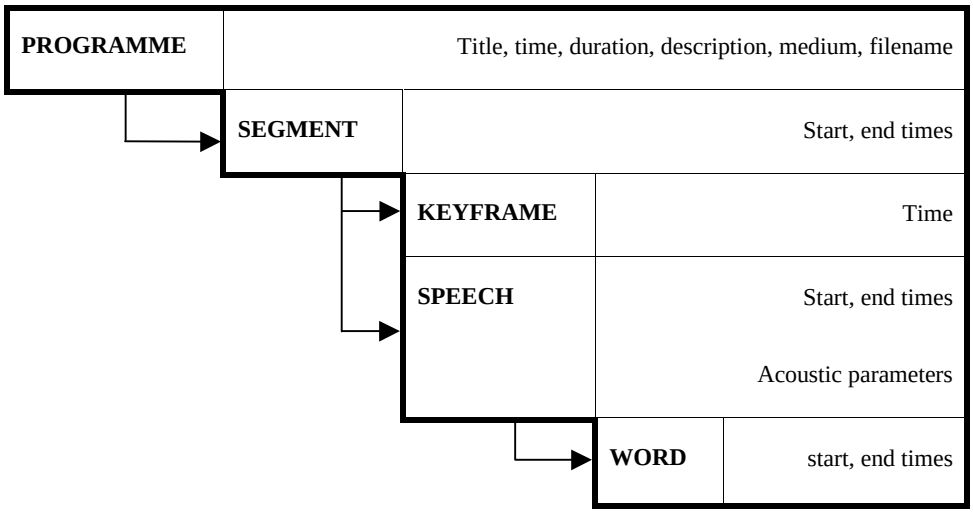

Figure 2. Structured view of programme content

Although possible, the XML documents are not stored directly in the database because XML attributes are untyped. Instead we create typed, indexed database objects corresponding to the XML nodes. Keyframe thumbnails, if any, are extracted from the MPEG file and stored in the database for rapid display. Thus all the content extracted from the MPEG stream is stored in the database. The database acts as a web server. Database objects are used to generate HTML that provides the searching and browsing interface to the system. All references are bi-directional, allowing rich navigation and browsing of the archive.

5. BROADCAST NEWS RETRIEVAL

The news retrieval system was intended to build a long-term archive of television and radio news, collecting programmes throughout the day and updating the index each night. Sources included BBC Radio 4's "Today" programme and the BBC TV 9 o'clock news.

The Entropic Truetalk Transcriber (Valtchev *et al*) was initially used as a speech recognition engine. This is a sophisticated speech recognition package capable of state-of-the-art performance on dictated, broadcast and conversational speech. Fifty million words of the British National Corpus supplemented with 2 million words of TV closed-captions (included twice for emphasis) were used as resources for language model training. A vocabulary of 60k words (with pronunciations from the BEEP lexicon) was defined and used to train a back-off trigram language model with relatively high cut-off values for compactness.

The acoustic training material used the second channel of the WSJCAM0 (Robinson *et al,* 1995) corpus with a perceptual linear prediction based parameterisation scheme. State-clustered triphone hidden Markov models (HMMs) were constructed. Various broadcast TV and radio shows were meticulously transcribed and used to source MLLR (Leggetter & Woodland, 1995) acoustic adaptation to the particular domain, improving the speed and accuracy of the transcription process.

Due to the potentially vast quantities of material requiring transcription each day, the recognition process is optimized for speed rather than accuracy. This is further justified on the basis that high recognition accuracy (though obviously desirable) is not crucial for information retrieval tasks. Consequently, recognition is performed in a single pass with reasonably tight pruning parameters in around seven times real-time on one 166 MHz processor. The speech recognition process easily lends itself to parallelism. Segments classified as containing speech are distributed between the processors and the output combined. The word error rate on a sample half-hour news programme using the adapted models was 30.7% compared to 37.6% without adaptation.

We experimented with a number of content extraction and retrieval techniques, summarised below. Our aim was to determine which techniques would prove to be at all useful for the task of news retrieval.

Keyframe retrieval. The DART project has also been investigating image retrieval. We applied some of our work in this area to the keyframes selected during the indexing process to represent segments visually. There tends to be a large number of scene changes hence keyframes in news programmes, and so we paid a high computational cost for keyframe indexing. We tried simple colour histogram techniques as well as more elaborate image-segmentation schemes. Keyframe comparisons may be useful for detecting repeated images such as the TV news anchor or repeated use of a video clip across a number of broadcasts. The histogram techniques seemed to work well for this purpose. While possible, more complex image searching is much slower and less useful than text IR for finding semantically related scenes across the entire archive.

Acoustic searching. The mean and covariance of the feature vectors across each segment are generated as a by-product of the audio segmentation scheme. This forms an "acoustic fingerprint" of the segment that lends itself well to acoustic searching based on a "find similar" paradigm. The acoustic properties are compared using the Kullback-Leibler (KL) distance (Siegler *et al*, 1997). This required a linear scan for comparison as the KL distance is non-metric; it does not satisfy the triangle inequality thus preventing the use of an efficient indexing structure such as M-Trees (Ciaccia *et al*, 1997). Despite this drawback, it was possible to find other instances of the same speaker (where these exist) to a surprisingly high degree of effectiveness. An obvious extension would be to use additional Gaussian mixture components to model the acoustics in more detail. This was avoided however to reduce computational cost in both indexing and search. Acoustic searching is particularly useful when searching for regular speakers, such as the presenter or reporter. For example, one could find previous reports on a particular story from a single report by the relevant correspondent.

Combined acoustic and keyword queries. The facility was added to retrieve items based on a score formed by combining acoustic similarity and text based searching. For reasons of efficiency, this involved post-processing the text based query results based on rank position using acoustic evidence. This feature suffered from the problem that no theoretically sound method of combining the two scores could be established. Furthermore, the feature was considered to be confusing and provided no clear advantage to users over either acoustic or keyword queries alone.

Lexical merging. As mentioned earlier, the segmentation scheme has an optional lexical merging step. A term vector is computed for each segment. Consecutive vectors are compared using a vector space cosine distance function to measure the similarity of the content of one segment with the next. Adjacent segments whose distance is below an empirically derived threshold are considered similar. These segments are merged. In practice, the lexical overlap between perceptually similar segments was often slight. For instance, a politician's answer might have little lexical overlap with the original question. An area for investigation might be the use of text segmentation schemes exploiting local context (Ponte & Croft, 1997).

OCR on video. Many news broadcasts contain scenes in which the name of the speaker or location is displayed on the screen in the form of a caption. This information is often not replicated in the soundtrack or closed-captions. We used thresholding and connected component grouping to detect

and locate possible text within video frames quickly. Regions containing text were then passed to the OCR engine. The quality of MPEG-1 video makes letter and word segmentation problematic due to blurring and DCT artifacts, so that conventional binary OCR fails at any threshold. We therefore developed a fixed-font greyscale OCR that avoids letter segmentation. Firstly, it correlates the image with each stored glyph at many horizontal positions, estimating the likelihood of that letter from the mean absolute difference (chosen to give robustness to noise and blurring). A set of likely letter positions is obtained. We then used dynamic programming to fit a horizontal sequence of letters, using rules about letter spacing and uppercase, lowercase, digit, punctuation and whitespace transitions. It uses a bigram spelling model derived from TV news teletext transcripts. A final "spelling clean-up" phase matched each word against misspellings of 60,000 known words. These were obtained by exhaustive application of common substitutions such as a↔o↔e, D↔O, G↔Q, rn↔m and i↔l↔f↔t. The system worked well on large captions consisting of capital letters such as news headlines. In this case, the word error rate was below 10% allowing conventional word-based information retrieval. Error rates rose sharply on smaller captions with lower-case letters (such as bylines) to around 25%. Non-words such as numbers and URLs were seldom transcribed correctly.

6. TV RETRIEVAL

Digital TV recording systems are now commercially available, including systems which attempt to learn those programmes which the user prefers to record. However, they still require some amount of VCR programming effort. We wished to assess the feasibility of recording all UK terrestrial TV output, thus ensuring that no programme is ever missed.

The TV recording system collects programmes throughout the day and adds them to the database on an hourly basis. Incremental indexing allows postings to be appended to the inverted index without requiring a complete rebuild of the index. Each night we purge week-old programmes and rebuild the text index. In future work we aim to achieve near instantaneous update of the database by processing the audio and video streams while they are being broadcast. This is not a requirement for a TV archive, but it is perhaps more important for a broadcast news service.

For high-quality speech recognition, it is important to have good acoustic and language models for the target domain. Constructing and maintaining such resources for general TV output would require considerable time and effort. Speech recognition is also computationally expensive compared to the other indexing stages. The news indexing process took four CPU hours for a half-hour broadcast. Therefore, to extend the system to a full range of television broadcasts on multiple channels we opted to use the closed-caption teletext service, when available, instead of ASR. The proportion of programmes that are captioned is regulated by law. For instance, in the U.S., the law requires that in the near future 95% of new programmes and 75% of archived programmes be captioned. The majority of programmes in Britain are currently captioned and we initially considered using ASR for the remainder. These programmes tended to fall into two general categories: live broadcasts and overnight programmes. In the first case, live broadcasts such as sports coverage tend to be transcribed by ASR poorly due to the nature of the material. In the second case, we found no demand existed to see this kind of programme. We therefore decided against using ASR in the day-to-day operation of the system. It is also worth noting that teletext subtitles are often a paraphrase of what is being said. Sometimes they add extra useful information, for example, indicating the presence and genre of music being played or a particular sound effect.

To a certain extent, achieving high retrieval effectiveness in the classical IR sense is not necessary to produce a useful and enjoyable system for our users, and we have made no attempt to measure it. It suffices to locate perhaps just one or two programmes or articles that the user might find of interest. Our measure of success is therefore not a recall-precision graph but a high usage from our colleagues. Usability is highly important; it must be simpler to use than programming a VCR.

The system presents the user with a TV guide, giving programme titles and descriptive information obtained from the teletext TV guide. The first sort of usage we observed was casual browsing of the

TV database via this list or a title search for a particular programme. Later people began to use within-programme browsing. The system presents a programme as a sequence of segments with associated keyframes and transcripts. Figure 3 shows the interface for browsing programme content. Each segment is presented with its associated keyframes and teletext or ASR transcript. The speech units (containing speech from a single speaker and environment) are marked with coloured bars below the keyframes. The programme browser is used to find particular parts of a broadcast, for example, the goals in a football match. These can then be played on demand.

As a supplement to the time-ordered programme guide, we added a simple classification of the programmes. The classes were chosen manually and included chat shows, news programmes and films. Selecting a class retrieves programmes by searching against the programme title. For example, news programmes are selected by searching the title for "news". Of course, there may well be news programmes whose title does not contain the word news, we are therefore evaluating the effectiveness of using programme content to classify programmes. The teletext programme guide also marks films with "FILM:" thus making their retrieval trivial.

In order to lessen user effort, we implemented a simple persistent query mechanism, using the archive as a TV filtering system. These persistent queries, expressed in terms of keywords to be found in programme content and/or titles, are stored in a user profile that can be edited within a browser. In addition to acting as a convenient shorthand for common queries, by including an email address in the profile, a user can be informed of the addition of relevant programmes. At the end of each day, users are emailed with links to programmes matching their queries from that day's broadcasts. This has proved to be a popular feature among our user community, who use it to track interests such as hobbies or sports. The system often recommends articles from magazine programmes. Since the majority of such programmes are irrelevant, the user would not ordinarily have chosen to watch them. Using the TV archive, only the relevant clips need be viewed.

Behind the TV archive lies conventional text retrieval. However users often think of their information need in terms of their interest, not of keywords which might be found in programmes about that topic. A typical example would be to search for *golf* rather than words more likely to occur in golf commentary such as *fairway*, *hole* or *birdie*. To help bridge this gap in terminology, when entering a TV interest (i.e. a persistent query) a list of suggested related terms is provided by Local Context Analysis (LCA) expansion (Xu, 1997). The LCA database consists of all previous TV transcripts, not just those for which the MPEGs are kept on-line. Clearly there is a great deal of work from the field of information filtering which could be used to build up a better interest profile by tracking which programmes the user found interesting (i.e. relevant).

The unit of programme retrieval is the segment. However, multiple segments from one programme may be returned. Some experimentation and user comments indicated that programmes with large numbers of low scoring segments were often more relevant than high scoring single segments. Retrieved segments that are consecutive also indicate a high relevance. We therefore group matching segments by programme, adding scores from each segment to give an overall score. This is a compromise between segment-based and programme-based ranking and further study is necessary in this area. This demonstrates a significant difference from clip-oriented broadcast retrieval systems, in which the retrieved segment ranks are the sole indicator of relevance.

7. CONCLUSIONS

We have described two systems for retrieval of broadcast television and radio. The motivation for developing the first system was primarily to develop our core technology in producing a speech recognition based news retrieval system. Using this as a foundation, the AT&TV system demonstrated the feasibility of large-scale digital archiving of broadcast media, using off-the-shelf hardware.

The media can be exploited in many ways. Image and acoustic analysis offer different ways of searching the material. Speech-recognised text and closed-caption transcripts are, however, by far the most useful and convenient means to search the data. While in some domains, such as education or news, the ability to search the collection is essential, in others it is of secondary importance to simple browsing of the media. Maintaining a profile of persistent user interest queries reduces the user effort required in tracking a particular topic. The capacity to inform users by email of relevant programmes was also successful.

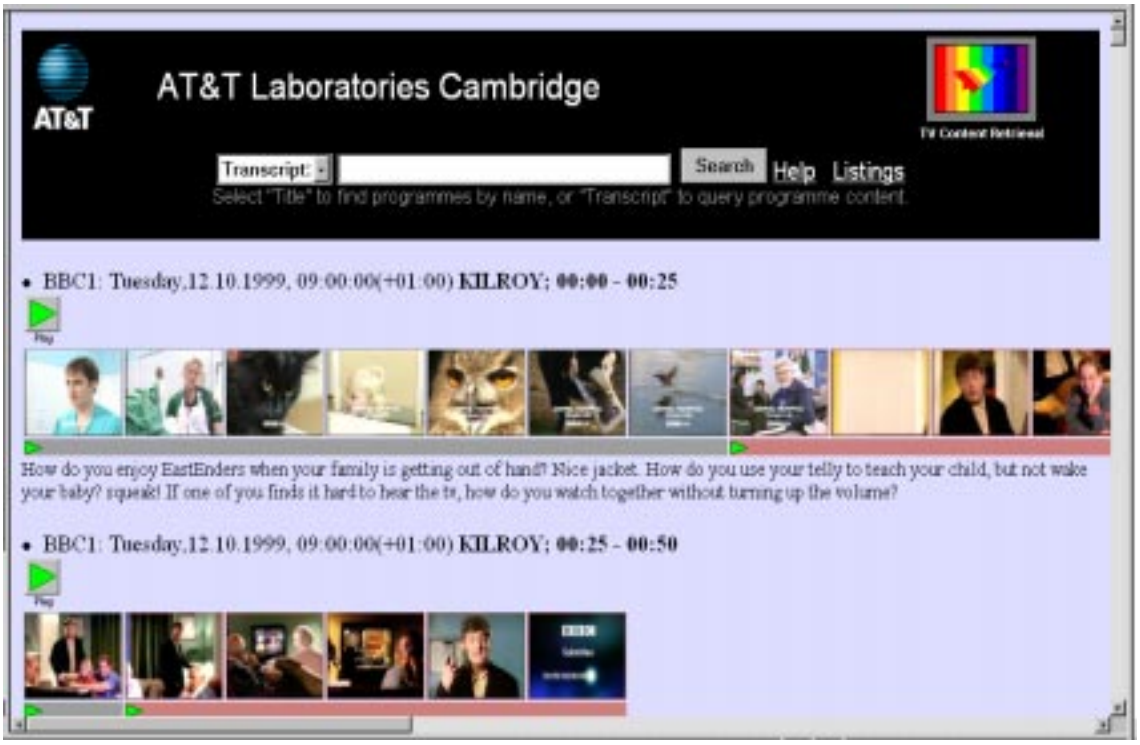

Figure 3. Browsing a programme, segment by segment

The TV retrieval system proved much more popular with our colleagues than the news system. For a general TV archive, aimed at entertainment, users just want to find something to watch with the minimum of effort. This can be achieved by simply providing a list of available programmes, perhaps organized into a number of classes by genre. Within-programme browsing is most useful for magazine-style programmes, yet it might ruin the enjoyment of a detective programme.

A current limiting factor of system usage we have found to date is the lack of broadband access to the home. Essentially, without resorting to efforts such as burning a VideoCD, this means that most people can only view programmes within the laboratory. We anticipate that these problems will disappear with the increasing popularity of DSL, cable modems and broadband radio systems.

REFERENCES

Brown, M.G., Foote, J.T., Jones, G.J.F., Sparck-Jones, K. & Young, S.J. (1995). Automatic Content-Based Retrieval of Broadcast News. *Proceedings of the Third ACM International Multimedia Conference*, San Francisco.

Ciaccia, P., Patella, M. & Zezula, P. (1997). M-tree: An Efficient Access Method for Similarity Search in Metric Spaces. *Proceedings of VLDB 1997*.

Gauch, S., Gauch, J. & Bouix, S. (1996). VISION: A Digital Video Library. *ACM Digital Libraries '96,* Bethesda, MD, 19-27

Gibbon, D. *et al* (1999). Browsing and Retrieval of Full Broadcast-Quality Video. *Packet Video,* NY, NY, April 25.

Jones, G. J. F., Foote, J.T., Sparck-Jones, K. & Young, S.J. (1996). Retrieving spoken documents by combining multiple index sources. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4-11.

Leggetter, C.J., & Woodland, P.C. (1995). Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models. *Computer Speech & Language,* Vol. 9.

Mills, T. J, Moody, K. & Rodden, K. (1997). Cobra: a new approach to IR system design. *Proceedings of RIAO'97,* pages 425-449.

Ponte, J. & Croft, W. B. (1997). Text Segmentation by Topic. *Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries,* pp. 113-125, UMass Computer Science Tech Report number TR97-18.

Pye, D., Hollinghurst, N.J., Mills, T.J. & Wood, K.R. (1998). Audio-visual segmentation for content based retrieval, *5th international conference on spoken language processing (ICSLP 98)*, Sydney, Australia.

Robinson, T., Fransen, J., Pye, D., Foote, J.T. & Reynolds, S. (1995). WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition. *ICASSP '95,* Detroit, USA.

Siegler, M., Jain, U., Raj, B. & Stern, R. (1997). Automatic Segmentation, Classification and Clustering of Broadcast News Audio, *Proceedings of DARPA Speech Recognition Workshop*, Chantilly, VA, 1997.

Turtle, H. (1991). Inference Networks for Document Retrieval. *Ph.D. dissertation*. CIIR, University of Massachusettes.

Valtchev, V., Kershaw, D. & Odell, J. (1997). The Truetalk Transcriber Book, version 1.0, Entropic.

Way, W., Glauert, T. & Hopper, A. (1994). Networked multimedia: The Medusa environment. *IEEE Multimedia*, 1(4):54-63.

Woodland, P.C. *et al* (1998). Improvements in Accuracy and Speed in the HTK Broadcast News Transcription System. *Eurospeech'98* , Budapest, Hungary.

Xu, J. (1997). Solving the Word Mismatch Problem through Automatic Text Analysis. *Ph.D. Dissertation*. CIIR, University of Massachusettes.

URLS

1. AT&T Laboratories Cambridge, *DART Project*. http://www.uk.research.att.com/dart

2. FasTV. http://www.fastv.com

3. Infomedia project. Publications available at http://www.infomedia.cs.cmu.edu

4. Media Site. http://www.mediasite.net

5. ReplayTV. http://www.replaytv.com

6. SpeechBot. http://speechbot.research.compaq.com

7. Telemedia Systems Ltd. http://www.telemedia.co.uk

8. Tivo. http://www.tivo.com