The Electronic Proceedings Project

Dr. Bob Jansen[1], Dr. Graham Barwell[2], Dr. John Robertson[3], David Hegarty[4], Peter Charuk[5], Greg Ferris[6] & Alan Walker[7]

1. Abstract

This paper presents interim results from a project to create an electronic proceedings of the AVCC Symposium on Australian Electronic Publishing, held in Sydney, Australia, in May 1996. The paper describes the architecture developed to support the electronic proceedings and the steps taken to produce the actual product, including end-user requirements, the data conversion processes, representation, interface and advanced retrieval issues.

---

[1] Chief Scientist, Turtle Lane Studios Pty Ltd, mailto: Dr. Bob Jansen <turtlelane@moreinfo.com.au>

[2] Senior Lecturer, University of Wollongong, Australia, mailto: graham_barwell@uow.edu.au

[3] Senior Researcher, CSIRO MIS, Australia, mailto: john.robertson@cmis.csiro.au

[4] Director, CADRE, Australia, mailto: d.hegarty@nepean.uws.edu.au

[5] Senior Lecturer, University of Western Sydney - Nepean, mailto: pcharuk@nepean.uws.edu.au

[6] Director, MSV Video Services, Australia, mailto: g.ferris@unsw.edu.au

[7] Alan Walker Indexer, mailto: alan.walker@s054.aone.net.au

2. Introduction

The use of the Internet and especially the World Wide Web (WWW) is gaining popularity for the electronic publishing of conventional publications, such as research reports, books, manuals, and so on. The majority of these products share a common feature: the use of 'electronic' paper.

With electronic paper, the reader is presented with a simulacrum of paper which allows concomitant manipulation. The terms used to refer to aspects of this simulacrum are inherited from paper: pages, headers, references, citations, margins, gutters, etc. Several technologies have been developed especially to mimic paper, such as Adobe's Portable Document Format (PDF), which presents readers with page structures manipulated with a pseudo hand. PDF has its roots in Adobe's Postscript technology and this explains its paper-based mentality.

Electronic paper does, however, offer one advancement over traditional paper, namely the anchor and link structure. Using this structure, an author or publisher may identify take-off points in the publication and the address of the target so that, at run time, the reader can associatively link to and show the target object. Unfortunately, current models only facilitate one-to-one anchor-link mapping, so that an anchor can only point to one target, a ridiculous limitation given what we are used to with paper. IntelliText (Jansen and Ferrer, 1997; Jansen and Bray, 1993) was a project which explored this limitation and provided true many-to-many mapping.

This paper presents interim results from a project that aims to move beyond the mediocrity of simulated paper and to realise the true potential of publishing in an electronic form. We discuss the results to-date of this project are discussed and consider the following matters: user requirements, architecture issues, conversion of the data, preparation of the transcripts; synchronisation of the various information streams; indexing of video content; delivery to a user; artistic participation in user navigation issues; and authoring of audio navigation structures. Wherever possible, we include examples of the product to highlight the issues. The paper finishes with topics for further research.

3. The Electronic Proceedings

In May 1996, the Australian Vice Chancellors' Committee Electronic Publishing Working Group[8] held a symposium in Sydney, Australia, to publish the results from the projects that had been funded by the Working Group to date, as well as to present conclusions

[8] Dr. Bob Jansen, one of the authors of this report and the project manager of the project, was a member of the AVCC Electronic Publishing Working Group.

realised by a number of specifically commissioned papers, authored by suitable experts, in key areas of fundamental concern to electronic publishing.

The symposium consisted of a series of presentations, with questions and answers if time permitted, followed by short updates to each of the commissioned papers[9], with the final session consisting of a lengthy discussion of points either raised during the day or identified by the Working Group prior to the symposium. Each of the first six presentations were of forty five minutes duration with the remainder limited to ten minutes per presentation.

The Working Group allocated A$20,000 to the preparation of the electronic proceedings, with the remainder being provided by the collaborators[10], who had a free hand in the design of the final product. From the start, it was decided that the project would not merely do an 'HTML job': run transcripts or submitted papers through a suitable converter and load it on a web site, but would try to present the *event*.

Any event consists of more than just the transcripts, or a video recording: the communication channels used by presenters are usually more extensive that those available to an author and if the experience of the event is to be replicated, then these channels must be captured and made to work together as they do in real life. This does not mean that a live event is intrinsically superior to a book: novels will almost certainly remain in a linear form for which paper excels, but symposia by their very nature are rarely taken linearly. Attendees drop in and out of them, speakers vary in quality, most of the real action happens in the breaks, networking abounds and could in fact be said to be the *raison d' être* for the event in the first place.

The symposium was duly recorded in video and audio and the task of the project was to convert this ten-hour dataset into a form for electronic publishing. In addition, the speaker support material was copied, most of which was already in electronic form as Powerpoint™ presentation files or sets of HTML files. Finally, several participants commented on the presentations and the issues raised. These were to become guided tours of the day's proceedings and were recorded in video and audio.

The initial tasks of the project were two-fold: to determine what type of material was required and then to convert the recorded material into electronic form. The former activity involved determining the user requirements of an electronic proceedings, articulating the vision of the team in terms of what they would like to do rather than what the technology constraints were, and then to develop an architecture that compromised the vision as little as possible.

---

[9] The commissioned papers were updated and published in conventional book form (AVCC, 1996) and made available on the web (http://www.adfa.oz.au/Epub/ summary.html).

[10] To-date, the project has cost in the vicinity of $150,000.

™ Powerpoint is a trademark of Microsoft Corporation.

As a guiding principle, given the sheer volume of available data, ie. 10 hours of moving image, 10 hours of audio, complete transcripts, all speaker-support material, six guided tours[11] and indexes, it was decided that the capabilities of desirable delivery systems would not constrain the design. Hence, if a desirable function required a system that could not, for example, be net-based, then the world wide web would be sacrificed in favour of functionality.

4. User Requirements

An early meeting of the development team developed an understanding of the user requirements. The requirements agreed to were as follows:

- Given that the data consisted of a number of different types/media, the user must be able to find chunks of relevant data irrespective of type.
- Once relevant chunks had been identified, the user could select, from a number of available channels which particular channels to engage.
- All channels of data must remain synchronised throughout playback.
- The user should be guided to the chunks of available information using state of the art indexes and revert to brute force searching only if necessary.

5. Architecture

The architecture of the system is fairly straight forward and is similar to the architecture of common video compilation tools, such as Adobe Premier; namely a set of parallel time-based channels of information intersecting with another set of time-based information representing the various guided tours but keyed to a different time-code.

[11] Due to time constraints, the guided tours were recorded but not processed further. It is hoped to complete this activity at a later date.

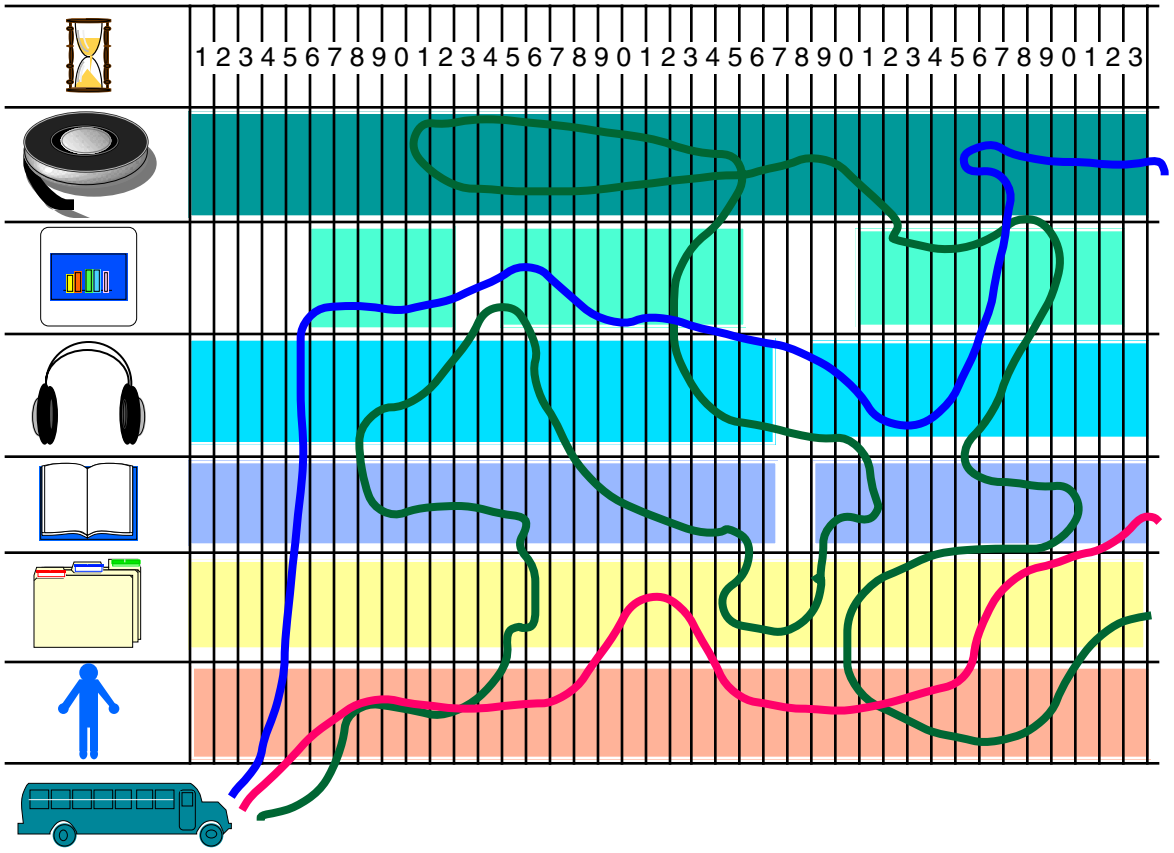

Figure 1. Architecture of the Electronic Proceedings

6. Conversion of the Data

The project has undertaken the mammoth task of converting the recorded information into the required formats for delivery. The analogue video tapes were digitised, after agreement as to a suitable codec, verbatim transcripts of each verbal interaction were created, and the completed transcripts were then edited (see the section on preparation below).

The enormity of this task should not be dismissed. Despite a broadcast quality recording, the transcriptions continue to contain errors which are time consuming to identify and correct. The digitisation of the recorded video and audio, being 20 hours in total length (10 video plus 10 audio), represents a major challenge to today's technology. The production of such a large amount of text synchronised with moving image is not yet supported by appropriate tools and must be manually accomplished (see the section on synchronisation below).

7. The Preparation of the Transcripts

The transcripts were prepared from audio cassette tapes of the day's events. A professional and competent transcriber keyed them in after which they were passed to a experienced editor for editing. A first draft of the edited transcript of the first session of

the symposium was then prepared and problem areas were identified. The project team determined transcription policy, whereby the transcriber was instructed to produce verbatim transcripts without correction, and agreed upon editorial policy. The first draft of session 1 was revised in the light of agreed editorial policy, then returned to the project team for final comment after which the remaining sessions were edited.

A major decision and one that caused weeks of, sometimes, heated discussion revolved around the point of having transcripts. It was agreed by all that transcripts were to be provided:

• So that users can check what was actually said by speakers at the conference;

• So that the hearing-impaired can see what was being said;

• So that people can read on screen as they listen to and/or watch the speaker; and

• So that the text can be used for indexing or searching.

The transcripts are not designed

• So that they can be used in print form for later reference; nor

• So that excerpts from a print version can be quoted elsewhere

In the formulation of these objectives it was necessary to first acknowledge the differences between speaking and listening, and reading silently. Spoken English is significantly different from written English; the spoken form of the language is accompanied by a number of non-verbal cues to meaning, such as tone, inflection, facial expression, etc., whereas the written form has to rely on the words themselves to make meaning. It is less flexible and more formal because it is fixed where the spoken form is ephemeral. A transcription which tries to record an oral performance and to do so in a way that makes it suitable for use as a printed document is attempting to meet significantly different and sometimes conflicting requirements. While accounts of parliamentary debates, for example, are printed records of oral performances, they are not accompanied by audio and video recordings and thus differ from the transcripts most appropriate in an electronic proceedings. The style of transcript more appropriate to the situation when a user is listening to the oral delivery is exemplified in the CD-ROM, *Long Time Olden Time No White People Bin Here* (Penrith, NSW: Firmware Design, 1993). This is an oral history of Aboriginal experience of life in the Northern Territory earlier this century in which the speakers' words appear on screen as they speak. Pinker (1994, pp224) has further discussion regarding this important topic.

If the transcribed text is to be read at the same time as the presentation is heard, then the transcribed text must be as accurate as possible in order to prevent readers being confused by the text not matching what they hear. In this case, apart from supplying punctuation and determining sentences and paragraphs, the editor should only ignore the kinds of sounds, *um, ah* and the like, which we normally do not pay attention to while listening. The sense of immediacy will be strong, even if the transcript is not ideal for silent reading.

This style of transcription will be suitable for those who wish to check on particular words, phrases or names which a speaker used. If the transcript is to be used as a printed text, then it will require different editing with much greater attention to the features which make it easier to read. The editor, for example, will have to revise the material to clarify the sentence structure or correct grammatical or other slips in the oral delivery.

The usual presumptions behind conventionally published conference proceedings are to make the material widely available, to preserve it, and to give it authority. The printed material frequently differs from the actual presentation; the content may have been revised by the presenter, but the printed paper will certainly be much more formal (layout, style, language) than the conference presentation. If users of the electronic proceedings want to print out the papers, parts of papers, or question and answer sessions, then they will want something which is designed to be read on paper. This means that a transcription which attempts to give a highly accurate rendition of what was said may be entirely unsuitable for printing out. In addition the paper presenters may well feel unhappy with an oral performance being fixed in print with all the features of oral delivery intact and lacking the conventions of printed papers (eg, referencing). Transcripts in an electronic proceedings therefore need to acknowledge that they not designed for the medium of print and that the usual presumptions about conference publication will be met in a different way in the electronic medium. The above objectives were thus appropriate for the project.

With the absence of written papers prepared by the presenters, an unexpectedly time-consuming aspect of the editorial work involved the checking of the proper names (eg. persons, products, places, etc.) and acronyms used by the various speakers. Indeed, once editorial policy had been determined, checking the transcripts against the audio cassettes and verifying the spelling of names and acronyms constituted the major part of the editorial effort in the project.

8. Synchronising the Various Information Streams

This activity investigated methods of synchronising the text data streams with the moving image and sound streams. The synchronisation of the moving image and sound streams themselves is a problem adequately addressed by existing technology.

We investigated various methods involving the use of Quicktime's™ text track feature for presenting superimposed text and moving image. Although this kind of superimposition is now technically feasible, it is more suitable for displaying limited quantities of text onto a moving image, as happens, for example, with sub-titling.

The text track feature of Quicktime was not suitable for this project as our text tracks contain 10 hours of transcriptions. Quicktime, although suitable for small amounts of text,

---

™ Quicktime is a trademark of Apple Computer, Inc.

had difficulty in retaining any notion of synchronisation: we were not after lip-synch, but synchronisation between what was being said and the superimposed text. Quicktime essentially failed to remain synchronised after a few minutes worth of text. The text fell behind the image and sound and then, in a flurry, tried to catch up again by streaming text across the image extremely rapidly, making the job of reading impossible. Once the text has become unsynchronised, we, as readers, enter the cognitive challenge of hearing one thing and reading another: a challenge best avoided for sanity's sake.

We also investigated several methods of superimposing text, including static text and streaming text like ticker tape across the image. The latter was totally unsuitable due to the cognitive problem of reading a moving data stream on top of a moving image, especially where the motion was in contrary directions. We found that the likelihood of losing one's place was unacceptably high and led to a confusion as to which track was the 'master' track and hence to be consulted as to one's location in the information space.

The superimposition of static text led to a number of problems in deciding how much text comprised a chunk - is it a phrase, a sentence, a paragraph? We found by trial and error that chunks were best delimited by pauses in the speech and that the chunk would remain on screen for the duration between the pauses. This does, however, lead to rapid chunk changing in the case of some speech structures, such as single names to introduce the next speaker, ie. "Bob" (with a rising inflection), or inadvertent slip-ups in the speech or the speaker's train of thought. A comparison with commercial sub-titling seemed to indicate that this chunking was the method also employed in this area but this is the subject of further exploration.

In the end, it was decided that the optimum technique was to have a separate text frame. This reduced the synchronisation problem in that a larger chunk of text would be visible and hence the synchronising was of lesser granularity. An advantage of this approach, realised only after building of the system was begun, was that the text could now be fast-forwarded or rewound, enabling the reader to anticipate what was going to be said or review what had been said. When the text is manipulated in this way, the text channel will re-synchronise at the next synch point.

9. Indexing of Video Content

9.1 Indexing processes

The following presents a list of activities performed in the creation of the index. It is provided, in terse form, to provide an indication of the processes involved and to place the resultant table in context.

9.1.1 Compiling table of contents

The first necessary step was to provide locators for the index terms. That is, the location of each topic on the video had to be logged. This was done simply by playing the video, noting the beginning and end of each topic of discussion, and recording the topics and time codes.

As a result of this process a table of contents for the video was produced. The table of contents refers to blocks of text similar to paragraphs in printed publications, which are larger than the 'chunks' referred to above.

9.1.2 Inputting table of contents

It was not practical to input the table of contents data while watching the video. Therefore, the details were hand-written first, then input. Some editing of the data was possible whilst inputting.

9.1.3 Compiling index to title level

The table of contents had then to be converted into the form of an index. That is, the topics had to be arranged into alphabetical order, headings and subheadings devised, and the correct locators (time codes) assigned to each index entry.

As a first step, an index was created with as much detail as contained in the titles of each presentation. The names of presenters and other contributors were included. This work was input directly, by reading and rearranging the table of contents file. As a result a short, general index, providing a general format and structure, and including some cross-references, was created.

9.1.4 Compiling index to detailed level

The extra detail included in the table of contents could now be incorporated. This expanded the index considerably. This was done manually, by consulting the printed table of contents, and adding details to a printout of the index. Many new cross-references were added at this stage, as synonymous terms and relationships between topics became apparent.

9.1.5 Inputting the index

Additions to the index were input. Following standard indexing practice, a record of cross-references was also kept.

9.1.6 Proofreading the index

The input was proofread against the hand-written draft and against the table of contents.

9.1.7 Editing the index

Two distinct editorial processes had to be undertaken.

First, the accuracy of the index entries had to be confirmed. A particular problem in this project was that it was not possible to do this from the source of data, which is audio-visual. Checking of spelling, particularly of the names of software, systems and people, must be done from printed sources. The availability of reliable printed sources will influence greatly the accuracy of the index, and the time taken to compile it.

Second, the index itself had to be edited for consistency and useability. The indexer has to ensure that references to the same topic are brought together and not scattered, that useful cross-references between related topics have been made, and that the terminology is succinct and clear.

9.2 Time needed for manual indexing

The time needed to index the proceedings manually bears a close relationship to the real time of the video. Therefore the various processes may be expressed as a percentage of real video time:

Indexing process	**Time Taken** (as percentage of real video time)	
Compiling table of contents	122	
Inputting table of contents	80	
Compiling index to title level	37	
Compiling index to detailed level	56	
Inputting index	63	
Proofreading index	17	
Editing index	50–100	(depending on availability of printed sources)
TOTAL	**425–475%**	

According to this experience, the manual indexing time for 6 hours of real video time would be 25.5–28.5 hours. The cost of this work, at the Australian Society of Indexers' current recommended rate of A$35 per hour, would be A$900–$1000 (in round figures).

If the indexer were working from a transcript, rather than (or as well as) from audio-visual sources, this time and cost would be shortened, by perhaps 30%.

9.3 Factors affecting indexing

This section is included to indicate how the next indexing job could be improved.

9.3.1 Textual backup

The lack of textual backup, as a guide to structure and content, requires the indexer to do more research and verification than is usual. Textual backup includes transcripts, list of participants and copies of overheads.

The lack of a view of overheads on the video will also adversely affect users' comprehension, since their content is not fully covered nor clearly stated by the audio or the transcript.

While it helps the indexer (and the transcriber) to have attended the seminar, the main problem is in proper names, eg. names of people, names of systems and acronyms for software. Accuracy in representing these in the transcript and the index is impossible when only audio-visual sources are available.

9.3.2 Audio quality

Inaudibility of some parts of presentations (again, especially acronyms and technical terms) is also a problem. This could lead to errors and inaccuracies in the transcript and the index. It will also affect users' comprehension. Some parts of presentations are mumbled, and there are quick or quiet asides, which users may not catch. Some questions and comments from the audience are inaudible, especially at the beginning of the seminar.

9.3.3 Range of locators

It is normal in indexing text to provide references to both the beginning and end of a section in which discussion occurs, such as a range of page numbers (eg. 34–39). Indexing to both the beginning and end of sections of the video has been done, but hot links to the video can only be made to the beginning of the relevant section, so that some information will be lost to users, unless they check back to the index.

9.3.4 Menus and contents lists

It was necessary in this project for the indexer to create a table of contents before compiling the index. The table of contents, which provides locators, is also used as a menu (with hierarchies), and thus as an alternative way into the content of the video.

9.4 Use of the manual index

When the manual index is incorporated in the electronic proceedings, its use is maximised if hot links are made:

a) between index entries and the location on the video (using time codes), and

b) between cross-references in the index (eg. *see..., see also...*) and the index headings or subheadings to which they refer.

9.5 The index of the Electronic Proceedings

The final index for the electronic proceedings consisted of a file of terms and associated timecodes indicating where in the video space the term occurred. This index was able to be used in two ways: as an alphabetical list of terms (a conventional index) and as a time-sequenced list of terms (similar to a detailed table of contents) enabling, at any time, the relevant index terms to be displayed.

The time-sequenced form has proved the most interesting to-date, providing the reader with an associative navigation facility whose content changes over time. It also enables readers to follow a particular theme without breaking their concentration in having to open the alphabetic index first.

10. Delivery to a User

This activity represented the most challenging aspect of this project. Current delivery methods are limited to the Internet/World Wide Web, CD-ROM/local hard disk and video tapes. The latter is included for completeness but was not considered as it represents traditional technology.

Of the remaining two, the CD-ROM/local hard disk represents the most powerful option provided a large enough disk can be made available. The Internet/World Wide Web is the most challenging.

After several attempts at serving this large information base, it became apparent that conventional HTML/video technology was incapable of performing the tasks required of it. A search of other technology revealed a new development in delivering multimedia material, including video, across the Internet. WebTheatre, from Vxtreme (http://www.vxtreme.com) provides a mechanism for delivering synchronised multimedia material based on the same model adopted by this project, parallel time-based channels but in this case adding a separate controlling channel to provide the control of what is displayed, where and when. The authoring environment, running only on PC-based equipment, compiles into standard HTML for delivery across different platforms. After testing, the project adopted WebTheatre for delivering its content. This decision necessitates the reader to install the WebTheatre client in their web browser but this was determined not to be so onerous as to make the product unviable.

In terms of the actual screen layout, the event proved the guiding force. The video and speaker support material needed to be in a special relation to each other as they were on the day: the speaker was on the left and below the support material from an attendee's perspective. This was especially necessary in this case as several speakers referred to the support material displayed on the screen behind them. Thus if not in this particular topology, the speaker might be looking at the top right of the screen rather than at the support material, or at the index or the transcript. This may seem a moot point but is important as it was part of the 'event' related to the hall layout.

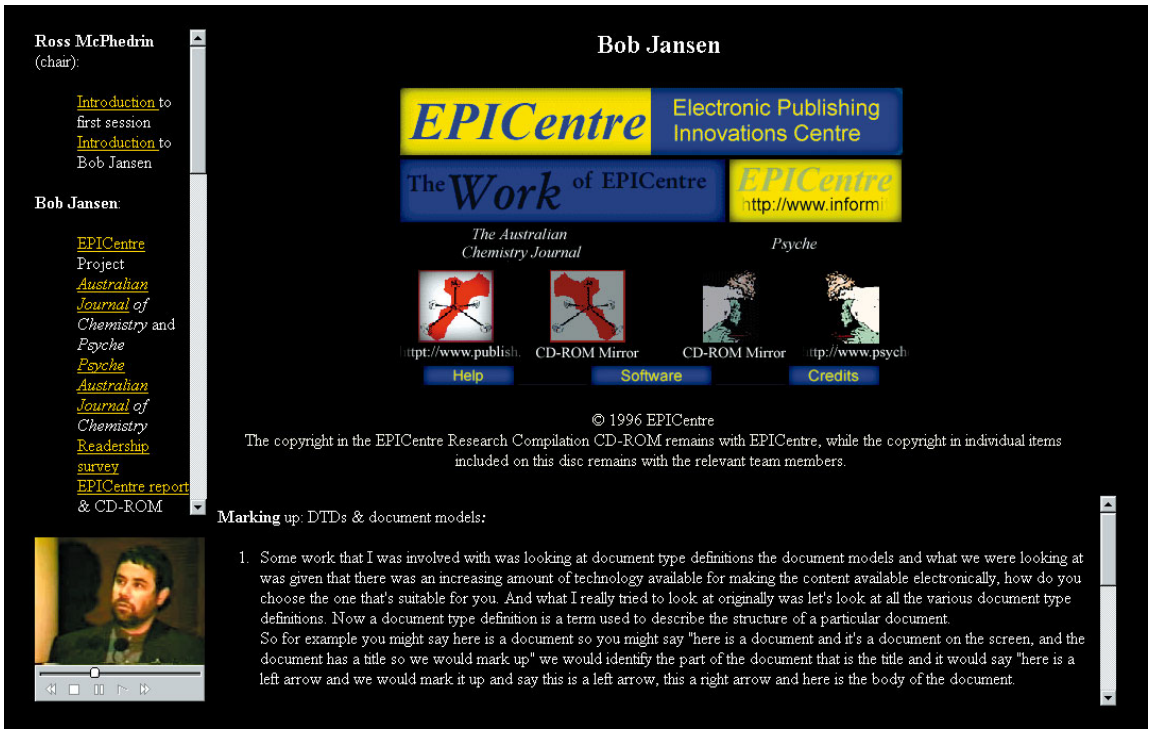

Figure 2. User Interface to the Electronic Proceedings. Note, top left is the time-based index, top right the speaker's support material, bottom left the video channel and bottom right the transcript. The audio is stored with the video but is engaged through the computer's speaker subsystem.

11. Artistic Participation in User Interface Issues

The brief for this part of the project was to investigate new ways of integrating different types of stored data, which, when translated into a digital format, gave an encapsulated experience of a one day conference: the artistic input.

The idea of an experiential map for the day's proceedings was the starting point. How does one develop a map of a day at a seminar? From one point of view, the basis for this came from the time that things occurred during the day and the space that they occupied within that time. The indexing process of the events was very important for the

development of the interface, since the index was aligned to the time codes of the videos of the event.

The interface design was a complex process because the idea was to encapsulate most of the activities of the day and to provide a simplified entry point to the gathered information. The interface needed to be simple, direct and visually enticing for the viewer to be interested in the information that was presented. The viewer needed to be able to find a relevant section quickly and seamlessly without much interruption to the thought processes.

The artistic input on this project has involved a number of research strands, particularly those to do with mapping, navigation and retrieving data from hyperspace. This has necessitated the investigation of other projects dealing with this amount and type of information. The key question for the experiential mapping of the day became what to leave out in terms of the map: a map being a simplified visual representation of complex information. This question led to what a participant in the conference or a might use as their preferred research methodology: an informed opinion looking directly for a particular strand of thinking, or a browsing methodology using memories of the day and not necessarily targeting specifics.

The data gathered shifts between

- Dense material - the reading of the papers, a visual representation of text and the person;

  to

- Loose material - the question and answer sessions and incidental asides, peculiarities that are key memories for some people but not all.

The user question became - what experience do I recall from this day or what amount of information do I require? How can I access it? Memory is another key element of the interface design. The seminar took place at a certain time, in a certain building and involved certain people. This could be represented visually through text and image that became the interface for the project.

The interface design being tested for the project is designed to capture the interest of viewers and hold their attention without them being bored or frustrated. Therefore, it was decided to incorporate a funky interface that was non-academic in feel to access the large amount of content and then be able to find the fine content.

This interface element also incorporated the notion of serendipity: a random entry to the information space.

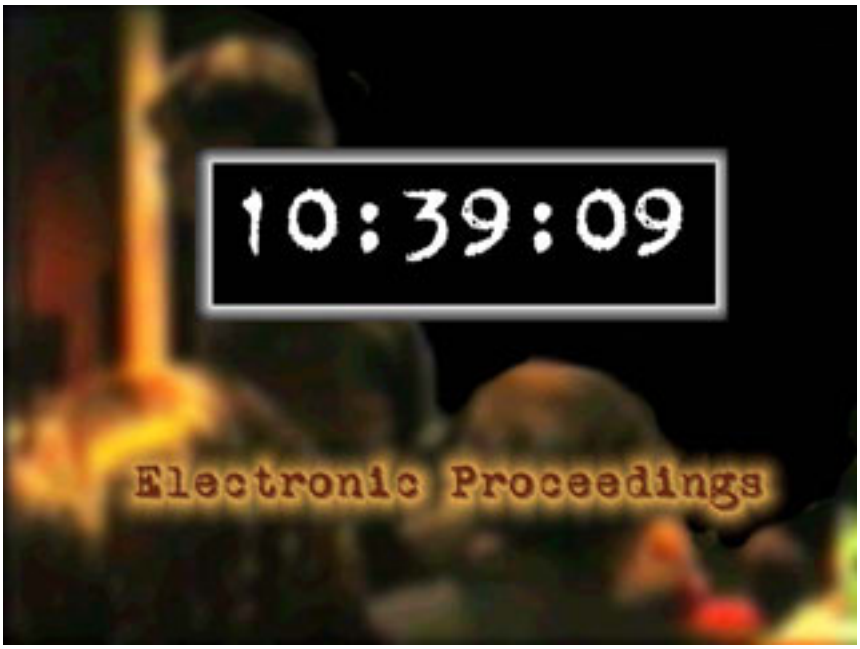

Figure 3. The Top-Level Interface

12. The Authoring of Audio Navigation Structures (AANS)

12.1 Background

There have been a number of research projects undertaken recently exploring the use of speech recognition technology for retrieving and filtering audio information (Kumar, 1995; Schauble; 1995; Wechsler, 1995). This sub-project, AANS, explored the use of speech recognition technology to produce a content base structure of video, audio and text information within a hypermedia delivery system. Specifically, can speech recognition technology be used to eliminate the need for hand crafted transcript or index development during the index and link generation phase of the hypermedia authoring process?

The hypothesis is that by using speech recognition technology a resulting reduction in the cost of indexing and linking text, audio, and video data will be realised.

12.2 The application problem

The production of large amounts of data containing human voice (film and video production, voice mail, radio, compact disks) has exacerbated an existing commercial dilemma; how to store this data in a manner which allows for efficient and economical reuse. Because of the high costs associated with human indexing of multimedia data and the lack of computer based support for this task, large amounts of audio data and audio/visual data are currently archived in an un-indexed state. This results in a situation where the media is very difficult to reuse and therefore is lost in practical terms.

12.3 Currently used indexing techniques

Indexing of sound data is typically done by hand. Human indexers create index terms or descriptors of the audio data as it is added to storage. The cost of this process can be very high, primarily as a result of the high level of human labour required by the process. Either the indexer must listen to the media to identify the contents or, if available, use a transcript of the audio data. The review by listening method is expensive because of the amount of time required for an indexer to audit the data. The alternative method, to index from a transcript saves the indexer time, but the production of the transcript can be extremely expensive - (Figure 4). Our experience with developing a transcript from ten hours of video was that this work required forty hours of a professional typist's time before the editor worked on it. Because of the high costs associated with these two approaches, much of the audio data stored currently and in the past is unindexed. Media owners are interested in developing techniques to reduce the cost of index development for existing and future media production.

One possible solution to this commercial impediment is to employ automatic speech recognition technology to support searching on audio media. We believe this approach has the potential to eliminate the need for transcript development and/or the need for indexers to review an entire audio monogram to determine the index concepts contained within. By allowing the indexer or end-user to search for occurrences of words within an audio track, a more cost-effective technique can be employed to retrieve and utilise media containing human speech.

The automatic speech recognition approach would be most effective with material containing speech data that conveys the semantic or pragmatic content of the video. For example, news or documentaries where the audio track is the principle conduit of information would be good candidates. Videos or film with little or no speech content, such as music videos or silent films, would not be useable by the proposed solution. Audio recordings or radio publications with little or no speech content would similarly be constrained.

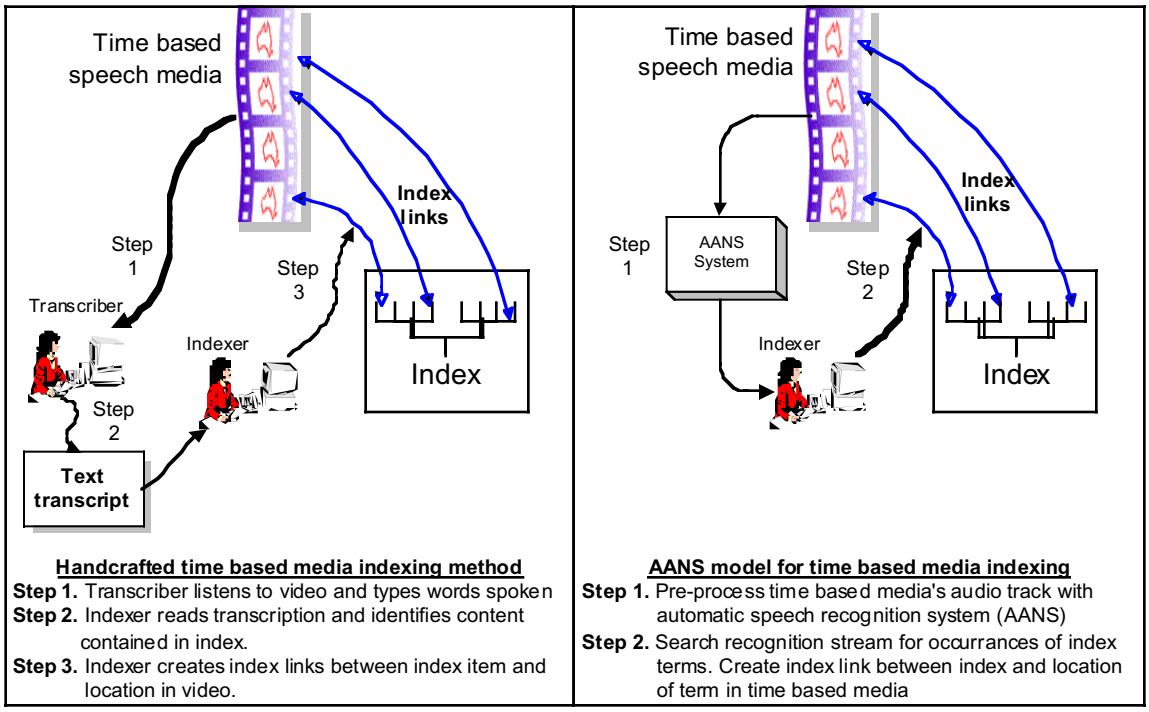

Figure 4. Indexing approaches for the time base media (with audio content)

12.4 Description of the technical problem

The ability to identify occurrences of words or phrases in a stream of speech data is called *wordspotting*. The objective of the wordspotting task is to identify the boundaries of a search term within a digitised continuous speech stream.

Because of the nature of the real world task being addressed - searching and indexing commercial multimedia archives - the audio data to be managed is particularly problematic for the current capabilities of automatic speech recognition technology. The wordspotting task is extremely difficult because it must manage data in a speaker independent manner, using an unrestricted vocabulary, and without training material from the archive itself. The automatic speech recognition system must be designed to manage:

- Speaker independence: Media archives typically contain data from more than one speaker. The automatic speech recognition solution must manage a diverse set of speakers. Members of this set of speakers can possess different vocal attributes, which will impact on the performance of the automatic speech recognition system. Differences such as gender, age, language and dialect variances introduce additional complexity. The AANS project video contained speech data from adult men and women with similar socio-economic backgrounds.
- Noisy Recording Environment: Because there can be no guarantee that the archive data has been recorded in a noise-free environment, the developed system must be able to manage unwanted background sound. The performance levels of most current speech

recognisors degrade significantly when environmental noise occurs (Hansen, 1996; Le Bouquin, 1996; Siohan, 1996).

Performance degradation is also caused by differences in the training and operating data's recording environments (Gong, 1995).The AANS data was recorded in an auditorium with ambient noise from the audience. The acoustics of the auditorium produced a low-level feedback on occasions. The Australian National Database of Spoken Language (ANDSOL) training data was produced under noise free conditions.

Another contributing factor to high signal-to-noise ratio (SNR) can be distortion introduced through the microphone and transmission equipment used (Acero, 1990). The distance between the speaker and the microphone can also introduce noise (Smolders, 1994). The speakers at the AVCC workshop moved about when talking, thus producing variance in the volume and quality of the recording.

- Open vocabulary: The research adopted an open vocabulary approach, which allows users to search on any term that might occur within the media database. The constraint of only being able to search on specific terms is too limiting for archives which are not thematically focused or when the development of a special thesaurus for the database isn't practical. Medium and large media archives commonly contain material that covers a broad range of topics. Also as new material is added, new topics are added. For this reason, the design took an open vocabulary approach.
- Continuous speech: The audio data will contain data spoken in a normal conversational manner. The recognisor must be able to manage continuous speech data.

12.5 Managing recognition errors when wordspotting

Given the extreme difficulty of the task being addressed: the location of occurrences of keyterms within a diverse quantity of audio data - current automatic speech recognition technologies can be expected to produce high error rates within the recognition output. Our research goal was to devise techniques to manage this error rate: to develop error management techniques which will result in commercially acceptable precision and recall outcomes. To this end the research focused on the trialing and evaluating the efficacy of approximate pattern matching techniques for keyterm searching within a recognition stream containing high error rates.

12.6 Preliminary results

Result analysis requires a common standard for measuring retrieval performance: recall and precision (Salton, 1989). Recall is the ratio of relevant occurrences of a search term identified divided by the total number of occurrences of the term within the audio track. Because we had the transcripts of the audio track, we were able to identify all occurrences of each test query easily. Precision is the ratio of the number of correct matching of the query term retrieved divided by the total number of strings retrieved. By evaluating the

precision and recall performance for a diverse set of search terms, an indication of the potential value of this approach can be identified.

The mean recall performance for this set of search terms was 54.72 and the mean precision performance was 31.74.

Because of their relationship between recall and precision performance (as precision performance improves recall performance usually declines and vice versa), the recall performance was necessarily reduced to ameliorate poor precision results. For all query terms 100% recall was attained, but precision performance at this level proved to be unacceptably low. A measurement developed by van Rijsbergen (1979) that combines recall and precision to find the optimum, balanced performance between the two measurements was used to identify each query term's optimum recall and precision balance. The reported 54.72 and 31.74 values were derived using van Rijsbergen's measurement.

13. Future Activity

This project has challenged each of us and our presuppositions in unforeseen ways. Yet, we have achieved something that transcends most of the electronic publishing seen on the Internet today. However, just as we have come far, we have a long way to go to realise our vision.

13.1 Delivery to users - continue the evaluation of WebTheatre/ Netshow[12]

The WebTheatre/Netshow development is an exciting development but one that requires extensive experience in melding the various components into a coherent whole. The software, as delivered, was not suitable for us, so we made major extensions by additional programming in JAVA. Much work needs to be done in this area, especially in tools to facilitate the loading of suitably formatted data into the authoring tool environment. These tools would ease the problem of identifying and marking up the synchronisation points.

13.2 Assemble the content from the various formats

In the existing project, the 'master' format was assumed to be the video recording. In other scenarios, this need not be the case and thus our authoring toolset must be capable of mastering from any format. This raises issues with the marking up of the various formats against the equivalent, in our project, of timecode. What is timecode in a non-video/audio environment or in an environment wherein the video timecode is outside the realm of the content, for example, a simulation over time?

---

[12] Since we began this project, the Webtheatre software and the company that produced it has been bought by Microsoft who are incorporating most of the components into Netshow version 3. Discussions with Microsoft seem to indicate most of our technology should work with some modifications, for example, the elegant Webtheatre authoring tool has disappeared to be replaced by a text editor.

13.3 Mapping interfaces

In the current proceedings, the index provides a major advancement in controlled navigation[13]. However, it is still difficult to ascertain where you are in the information space and what else may be of interest. Map-based interfaces should assist in this area and they are already targeted for the next phase of the research. Map-based interfaces might, for example, visualise an information space using topographic techniques to demonstrate the organisation of the information space and the relationships between individual information chunks. This type of interface would assist not only the reader but also the maintainer in determining where a new information object should go, by analysing its topography in the content of the information space, and how it relates to existing information objects, thus indicating what links need to be established.

13.4 Channel management

One area in which we are already progressing is channel management. Our vision is of an architecture that enables the reader to select which channels they would like to view at any time and to ensure synchronisation of the information. At present, this would require all the data to be downloaded to the client workstation, even that information not to be viewed. Server-side synchronisation is what is required, but how this interacts with available technology is currently unclear.

13.5 Audio Indexing

As indicated above, our research into audio indexing has really only just begun. We are planning a more concerted and better resourced foray into this area but recognise that a commercially-applicable result is still a long way off.

14. Acknowledgments

This project would not have been possible without the generous contributions of the following organisations: Turtle Lane Studios Pty Ltd, University of Wollongong, CSIRO, Centre for Applied Design Research and Educations (CADRE), University of Western Sydney - Nepean, MSV Video Services and Alan Walker, Indexer, who made available the people and resources to undertake the project for no direct financial return.

This project also acknowledges the financial sponsorship of the Australian Vice-Chancellors' Committee Electronic Publishing Working Group who provided the seed funding to realise the vision.

---

[13] See Jansen and Bray (1993) and Jansen and Ferrer (1997) for discussions regarding the Path knowledge type to control hypermedia navigation.

15. References

Acero A, *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Ph.D. thesis, Carnegie Mellon University, 1990

AVCC, *Key Issues in Australian Electronic Publishing*, Australian Vice Chancellors' Committee (pub), 1996

Gong Y, *Speech Recognition in Noisy Environments: A Survey*, Speech Communication 16(3) pp. 261-291, 1995

Hansen J, *Analysis And Compensation of Speech Under Stress and Noise for Environmental Robustness in Speech Recognition*, Speech Communication 20(1-2) pp. 151-173, 1996

Jansen B & Bray G, *Context & Knowledge Types Vs Serendipity*, in Proceedings of the 13th International Conference, Artificial Intelligence, Expert Systems, Natural Language, Avignon, France, 1993, pp85-95

Jansen B & Ferrer D, *IntelliText: An Environment for Electronic Manuscripts*, in Intelligent Environments, Droege P (ed), Elsevier Scientific (pub), 1997 (also presented at the 4R's Conference, Canberra, Australia, March 1994)

Kumar P, Narasimhalu A & Phanendra G, *Create-Time Indexing for Digital Video* Technical Report TR95-191-0. Institute of System Science, National University of Singapore, 1995

Le Bouquin Jeannès R, Faucon G & Ayad B, *How To Improve Acoustic Echo and Noise Cancelling Using a Single Talk Detector*, Speech Communication 20(3-4) pp. 191-202, 1996

Pinker S, *The Language Instinct*, Penguin, 1994

Robertson J, Wong YW, Chung C & Kim D, *The Use Of Approximate String Matching Techniques To Improve Audio Indexing Performance*, Internal Report CSIRO Mathematics and Information Sciences. Sydney Australia (currently unpublished), 1998

Robertson J*, A Hypermedia Authoring Methodology for the Reduction of Associative Link Errors*, Ph.D. Thesis University of Technology, Sydney, 1996

Salton G, *Automatic Text Processing: the Transformation, Analysis and Retrieval of Information by Computer*, Addison-Wesley Series in Computer Science, Addison-Wesley Publishing Co., 1989

Schauble P & Wechsler M, *First Experiences with a System for Content Based Retrieval of Information from Speech Recording*, IJCAI Workshop: Intelligent Multimedia Information Retrieval, MIRO'95, 1995

Siohan O, Gong Y & Haton J, *Comparative Experiments of Several Adaptation Approaches to Noisy Speech Recognition Using Stochastic Trajectory Models*, Speech Communication 18(4) pp. 335-352, 1996

Smolders J, Clase T, Sablonl O & Van Compernolle D, *On the Importance of the Microphone Position for Speech Recognition in the Car*, Proceedings ICASSP, Volume I, pp. 429-432, 1994

van Rijsbergen CJ, *Information Retrieval*, second edition, London: Butterworths, 1979

Wechsler M & Schauble P, *Speech Retrieval Based on Automatic Indexing*, Final Workshop on Multimedia Information Retrieval MIRO'95, 1995