BUILDING INTELLIGENT MULTIMEDIA INTERFACES

Michael Wilson
SERC Rutherford Appleton Laboratory
Chilton, Didcot, Oxon OX11 0QX, UK
mdw@uk.ac.rl.inf

Abstract

Hypermedia provides current authors with a new way to combine different media into single artefacts. These can be developed and marketed in the style of conventional media publications. Multimedia technologies not only bring the presentation of data stored as images, sound and video but also allow the incorporation of media specific input modes (e.g. natural language, speech, pen gesture), and the generation of images, sound and video from more abstract formats. Systems combining multiple input and output modes through an abstract representation are still a long way from the market, but parts of them (such as rules for automatically creating business graphics in spreadsheets) are being incorporated into existing products now. An example multimodal system (MMI^2) is described to illustrate the technologies which could be brought to the market independently. A more direct approach to introducing intelligence into multimedia systems is to incorporate it into conventional hypermedia tools. The MIPS multimedia presentation system which combines the presentation of an open pre-authored hypermedia network stored in the HyTime standard format with dynamically created web nodes containing answers to conventional database queries is described. This illustrates how conventional hypermedia tools can be extended to include intelligent automatic generation of multimedia presentations from retrieved data.

Keywords: Multimodal User Interface, Multimedia, Intelligent User Interface, automatic presentation

INTRODUCTION

The current interest in multimedia appears to have been sparked by the availability of sound and video technology on personal computers. However, these media have been available on supercomputers, mainframes and specialised workstations for some time. Similarly, personal computers are already expected to provide graphical user interfaces which include still images, menus, and the direct manipulation of graphics. Further, Personal Digital Assistants (PDA) are starting to use speech and handwriting based technologies which are not always considered within the scope of multimedia. Behind the use of these varied input and output technologies remain issues of appropriate storage formats, and the structure of application programs. This chapter argues for the introduction of dialogue control mechanisms and an awareness of context in application programs to support the automatic generation of multimedia presentations supporting the appropriate input and output media for information in the context of individual task performance.

Rather than progressing from the current state of the market towards the vision of intelligent multimedia systems, or starting with the vision and working towards the reality of the current market, a more dialectic structure will be used in this chapter. Firstly, contrasting visions of intelligent active computers and passive navigable information stores are presented. In contrast to these visions which have acted to drive research agendas over the last 20 years, the current practical state of the market is considered to provide a starting point for technological improvement. Next, a research demonstrator on the way towards the vision of an intelligent active computer is described - MMI^2. This is a long way from practical marketing but shows some of the issues being addressed by research which will filter into the market in small niches. Finally a more practical research prototype which lies between the MMI^2 research demonstrator and the current market is described to show how some intelligence can be introduced into multimedia systems to improve the portability of information without jumping ahead of the market's expected investment - MIPS.

CONTRASTING VISIONS

Developments in computer systems are often driven by visions. Two such contrasting visions are the anthropomorphic computers and robots publicised by Arthur C Clarke, and the Hypermedia vision found in Vannevar Bush's Memex system (Bush, 1945) or Ted Nelson's dream machines (Nelson, 1988). The HAL computer described in Arthur C Clarke's novel 2001 is an machine with human like intelligence and dialogue capabilities. Such a machine would result from artificial intelligence research modelling human reasoning and natural language processing. In contrast, the MEMEX of Vannevar Bush, is a large data repository where knowledge is used to cross index the information, but the access and search is performed under direct user control. The distinction between these views is whether the system includes any intelligence itself in the way it handles information, or whether data are treated as artefacts which humans are empowered to manipulate using their own intelligence and dumb tools.

When humans view images, listen to music or conversations, scan text or watch videos they analyse the data presented in these different media to extract information from them. This information is an abstraction of the data which can be combined with pre-existing knowledge and plans to support reasoning and decision making so that actions can be taken in the world and tasks performed. For different tasks, different reasoning processes will be used and therefore different information is required. The result of the analysis and the starting point for action for the human is an abstract meaning representation of the information required for a task. An anthropomorphic computer which performed analysis on line would have to include such a meaning representation to support reasoning, and have to select the appropriate information from the data for the task in hand (an analysis of the information required by different tasks is presented in Tansley & Hayball, 1993). Although the vision of computing based on a human-human conversational model is clearly anthropomorphic, it does not encompass all human attributes, and includes many non-human ones (e.g. Nickerson, 1977)

The alternative distribution of processing for the task is to separate the encoding and partial analysis of media into one task, the storage and distribution of media into a second task, and the reading of the distributed media, and further analysis into a second task. The first task performed by the author of the information includes sufficient analysis to support the further analysis in the third task performed by the reader, and user of the information. The responsibility for understanding the reader's final task, and selecting the information which is appropriate for it, rests with the author in the first task. The computer system performs the second storage and distribution task, and merely provides tools to support the author and reader in their two tasks.

Current developments under the heading of multimedia mainly follow the second of these two scenarios, being concerned with providing tools for authoring, storage and distribution, and reading of media. This chapter focuses on systems which follow the first of these scenarios involving abstract meaning representation with multi-modes of both input and output automatically selected to be appropriate to the task context.

THE CURRENT MULTIMEDIA MARKET

The multimedia market is currently very fragmented. Most market surveys suggest that training systems are currently the largest market sector, with point of information kiosks and games being the other main sectors (99% between them in 1991). Predictions suggest that the shape of this market will change with these sectors becoming saturated or at least stable in size. Data access and presentation applications are expected to grow to become the majority of the multimedia market by 1997 (Templeton, 1993). Communication companies such as AT&T are buying equity stakes in cable TV companies to ensure control of distribution, and in media production and asset owning companies (such as Hollywood studios) to ensure a supply of assets to distribute to domestic consumers. Computer developer companies such as Miscrosoft have made alliances with domestic equipment providers such as General Instrument and computer hardware developers such as Intel to integrate computer operating systems with domestic television systems (Microsoft have announced that Windows for TV will be launched in early 1994). Microsoft have also entered into agreements with fax machine, telephone and photocopier producers in the expectation of adding touch screens supporting multimedia to these domestic devices. Along with these moves to introduce multimedia information technology into domestic devices, market researcher Dataquest has suggested that domestic PC sales themselves will become 24% of the European market (30% of the UK market) in 1993, reducing the dominance of the business market. In addition to data access and presentation, the integration of computing and communications technology will give rise to a communication section of the market per se, for videoconferencing, videophone, voice or video mail (this will not be addressed in this paper, but is considered in others in this book).

Most published market predictions tend to accelerate the rate of market growth, partly because they are compiled from producers whose interests it is in to promote activity in a sector. In 1992 these suggested that the multimedia technology would be introduced into the market by 1993, with early entrants and technologists starting market growth by 1994, when bulk consumer growth would take off, so the multimedia market would rise to somewhere between $8B and $24B by 1997. These stages are certainly delayed since many producers cannot see a way to initiate large growth in the market. Products such as CDTV, CD-I, and Sony Bookman have all failed to capture the public's imagination by providing apparent desired benefits. Even some of the software video tools produced by major manufacturers have not provided sufficient quality of image for a market accustomed to TV or cinema quality.

Despite these setbacks there have been several marketing strategies envisaged for both the home and office, driven by games, desktop publishing, or voice mail. Games are moving up in complexity, with consoles moving from 8 bit to 64 bit by the end of 1993. CD-ROM drives are now required for many of the best selling titles, which may support their introduction into the home. CD-ROM versions of photographic libraries and video libraries have become available for desktop publishing systems which may introduce CD-ROMs into the office. Sound cards are acquired bundled with these drives, or explicitly for voice mail or sound annotation. Once the hardware is in place, and the viewing tools are incorporated into operating systems, then offices and homes are provided with multimedia PC's. Photo-CD will allow companies to establish libraries of their own images for incorporation into documents, and for home users to become accustomed to viewing their own images on screens. Once the hardware and software has an installed base then the market for individual multimedia titles can grow more easily.

Current multimedia training, kiosk and games products are Hypermedia artefacts which are discrete published entities. These may be produced as CD-ROMs for PCs, for CDTV players, for Sony Data Discman or any of several other reading platforms. In all of these cases though, they are single published artefacts whose production, marketing and sales follows the conventional paper or video publisher's product life cycle, and not a computer software one including maintenance and updated versions. The technology for producing these artefacts moves the control of their production from computer specialists to those accustomed to other forms of publishing. The authoring tools market will be initially large while empowering these creators, but will then diminish in importance as titles are produced. The estimates on cost of production of a hypermedia CD are in terms similar to the production of a one hour television program (about $100,000). The production teams are composed of similar individuals including cameramen, sound recorders, editors, directors, script writers and graphic artists (a recent book by Cotton and Oliver provides an impressively presented collection of images from available hypermedia systems collected by graphic designers and published by an art rather than a computer publisher - Cotton and Oliver, 1993). From the artists, typographers, or video editors perspective they are being provided with a new medium. Hypermedia changes the means of production since a single artist or designer can now sit at a single workstation and on that one machine orchestrate the complete span of media. It is possible to move seamlessly from typography to animation to illustration to image scanning or video editing to sound mixing, and at the same machine produce an entire interactive programme ready to be mastered and stamped on a CD-ROM, or networked to other machines. As the installed base of multimedia PC's in the home and office is established, the market for this class of hypermedia document will become established.

There are still considerable problems associated with this publishing metaphor for hypermedia production, mainly associated with competing distribution formats and copyright. Paper publishers are accustomed to a single form of paper publishing standard with variation in natural language Current Multimedia authoring tools produce their own proprietary formats, and can be distributed in several formats for different presentation tools. This confusing situation is many times worse than the competition between VHS and Betamax video standards which is the closest comparison available to most producers and publishers. A generalisation of Apple's QuickTime which is SGML compatible, HyTime became an ISO standard for hypermedia interchange in April 1992. It is expected that existing authoring tools will provide translators from their own formats into HyTime to facilitate portability, although none are yet available several major manufacturers have expressed support for the standard

(Newcombe, 1991; ISO, 1992). At a higher level, Kaleida Labs (a joint venture of Apple and IBM) have demonstrated a device independent multimedia programming language ScriptX on the way to producing a standard authoring system that will work on any computer.

Copyright and IPR issues associated with Multimedia products fall between those of computers and conventional publishing. Publishers, authors, photographers and other producers expect to retain copyright and gain a fee when library texts or images are used . Therefore they need to secure the assets and monitor access. The computer community is more familiar with buying software and then using it as they wish without paying by use. The conflict between these models and the legal resolution of them are addressed elsewhere (Lyons, 1991; Haynes, 1991; McIntosh, 1991).

In parallel with the introduction of various multimedia technology into the market has been the introduction of multimodal input mechanisms. Direct manipulation using a mouse or pointer is established as an input technique which corresponds with static graphic images or animation displays. Speech input techniques are often incorporate with text to speech output and non-speech audio output hardware and software. The accuracy in recognition reported each year at the workshops of the ARPA Speech and Spoken Language Program improve steadily, so that speaker dependant speech to text (about 95% accuracy on 20,000 word vocabularies) and individual word speech input systems are available for most major operating system GUI (e.g. Apple's PlainTalk) user interfaces as toys and are being seriously incorporated into personal digital assistants (PDA) for their specialist interfaces for limited tasks (99% accurate on 1000 word vocabularies). The main input mode for PDA's is a pen interface supporting handwriting recognition and 'Jot' based pen data. Jot is a data-interface industry standard, not for bit-maps, but for the position, colour and direction of strokes as well as timing and pressure information. Jot can be used to store images drawn using pens, but also signatures, and handwriting. Current cursive and non-cursive handwriting recognition systems are frustratingly inefficient (even a 99% recognition rate would fail on 22 letters in this paragraph with the result that one word per line would have to be re-entered). However, now that there are vehicles which use the technology for specific tasks, research is expected to improve the accuracy by both tailoring the recognition algorithms to the task as well as improving the base algorithms. Industry market predictions suggest that all lap top computers by 1997 will include pen interfaces. Further, the industry predicts that voice and pen technologies will combine together to fit the appropriate input mode to different tasks, as images, text, video or sound are chosen as appropriate output media. For example, while editing on the screen you might say "move this sentence [indicating what 'this sentence' refers to by simultaneously circling the sentence] to the beginning of this paragraph [simultaneously circling the paragraph] and change the formula to read as follows [writing out the equation]". Half small personal computers sold are expected to include both voice and pen interfaces by 1998 (Crane & Rtischev, 1993).

MULTIMODAL AND MULTIMEDIA SYSTEMS

Contrasting visions have been described of the intelligent assistant to task performance which can select information for a task from data, and communicate it to users in the most appropriate medium, and alternatively the information storage system where the information selection, and medium selection are performed by authors, while the end user is only a reader of that information. The current state of the market has also been described, showing how tools for information authoring, storage and presentation are becoming available, and how different input modes are being combined with these.

This section describes a framework for classifying multimodal and multimedia systems using these different input and output modes, and showing the other variations which arise from their use. The distinction between multimedia and multimodal interfaces is not obvious. Some authors regard multimedia as different presentation media and multimodal as different user input modes. Others make a distinction between the simple media which convey a message (e.g. video, sound, image) an the human sensory modalities which perceive it (e.g. auditory, visual, tactile). An important distinction in development is that multimodal systems are designed to be co-operative interfaces which actively choose the most effective and efficient presentation mechanisms for a user; whereas multimedia systems present the information in the medium which the author has provided. The most comprehensive classification has been produced by Coutaz, 1993, illustrated in figure 1. This describes a 6 dimensional space in which multimodal and multimedia can be categorised from the technical perspective of the system, rather than psychologically from the user's. The first dimension is that of the number of communication channels through which types of information can pass from human to system and back again. The second dimension is that of the direction of these channels. The remaining four dimensions categorise aspects of the rendering or interpretation of information.

The third dimension classifies the levels of abstraction available for each input/output channel. If each channel were to only present information encoded in a form explicitly for that channel (e.g. prestored encoded video) then it would only use a raw encoding. Similarly, speech input may only be recorded as a signal in a voice annotation system, it may be described as a sequence of phonemes, identified as matching a particular command function, or interpreted through a natural language module into a semantic representation including speech acts. The more abstract the representation used, or the more abstraction is included in the range of representations used, then the more power the mode is categorised as having. Information represented abstractly can be presented to the user through presentation mechanisms which use the mode which is most effective and efficient for conveying it. Information which has a raw encoding can only be presented through a mechanism specific to that encoding.

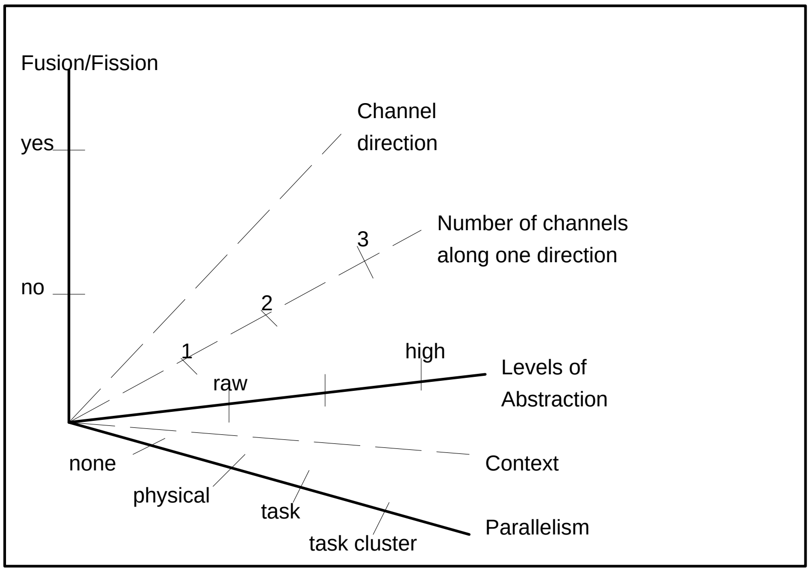

Figure 1: The MSM Framework: A 6-D space to characterise multi-sensory-motor interactive systems (After Coutaz et al., 1993)

The fourth dimension is that of context. This is not clear to interpret, although an example illustrates how the interpretation and abstraction of an utterance can depend on context. For example, in vi, when in command mode, typed text is interpreted as a command function, whereas in input mode the same text would be recorded without interpretation, abstraction or transformation. Contextual variables are used to control the abstraction process. This becomes more complex when dialogue, task and user models can be used to structure the context and its effect on interpretation of input and construction of output.

The context provided by dialogue task and user models will be used to interpret user input and to choose how to convey system output (e.g. Managers are a class of user who perform planning tasks; planning tasks require trend information; bar charts are a more effective presentation of trend information than a table of exact numbers; therefore to this user, in this task, a bar chart could be the chosen presentation form).

The fifth dimension is that of the fission & fusion of different modes. This dimension incorporates a binary distinction as to whether independent media can be combined together into single utterances to convey the intended meaning of the system (fission of a meaning between system output media) and a users input to the system can be expressed across more than one mode which are then combined together to form a single expression (fusion of different user input media).

The sixth and final dimension is that of parallelism. Temporal issues within modes such as video and sound, and between different modes are clearly important, and the ability for a system to interpret or render them in parallel is essential for successful fission and fusion of modes. Parallelism can occur at different granularities within the user interface: at the physical, task or task cluster levels. The physical level could be represented by the keystrokes that constitute a command, the task would be the command's function, and the task cluster would be the various commands and arguments entered to edit a document. Complementary levels of communication exist in system output where multiple media channels may be used in parallel.

MMI^2

To illustrate the architecture and knowledge required for multimodal interfaces, an example system will be described. Although other demonstrators (e.g. Feiner & McKeown, 1991) are more impressive in generating combined multimedia, they do not include advanced dialogue or input modes. The MMI^2 system was developed with the purpose of demonstrating the architecture and development method required to produce large scale co-operative interfaces to KBS (Binot et al, 1990). The first demonstration task used in this system is that of designing local and wide area computer networks for institutions such as hospitals or universities (figure 2 shows a screen from this system). A second demonstrator task was used to evaluate the generality of the architecture and the portability of the knowledge: the monitoring of local and wide area computer network performance (figure 4 shows a screen from this system). The overall architecture of the MMI^2 system is shown in Figure 3.

Figure 2: An example screen from the first MMI2 demonstrator showing different interaction modes with the underlying application:

The freedom provided to users by multimodal systems firstly relies upon the use of a abstract meaning representation common to all information sent to or received by each mode. The representation used for this must be able to express all such information in order to allow the choice of the most appropriate mode. In MMI^2 the language is called the Common Meaning Representation (CMR) which is a first order logic with extensions. This language is used to pass between the mode and dialogue management layers of the architecture, allowing a clear interface where different modes can realise (generate images, language, etc.) any CMR description. This is inefficient at encoding bulky media so it contains the logical information to be presented. The image, video or sound can then be selected to present this information. In the first demonstrator the exact geographical building structure as well as the logical structure of computer networks was encoded in this representation (as shown in figure 2), with a resulting slowing in performance. In the second demonstrator, map information (as shown in figure 4) was not encoded in CMR, merely the logical label of the map.

Each mode in the mode layer of MMI^2 has a generator to produce the mode's output from system generated CMR and a parser to produce CMR from user input. The modes supported are English, French and Spanish natural language, command language, audio, graphics for CAD diagrams, business graphics (charts, tables, pie charts, hierarchies), with direct manipulation by the user on these, and pen based gesture on these and the text modes. The natural language modes use conventional natural language processing techniques, the graphics mode uses explicit knowledge about the design of graphic presentations to produce effective and efficient presentations (Chappel & Wilson, 1993; after Mackinlay, 1986).

The second necessity for a multimodal system is that there is a common reference context for all objects. MMI^2 contains a Context Expert which stores all objects referred to in the dialogue and which provides the Dialogue Manager with candidates to resolve diexis and anaphora. Therefore each mode can refer to objects mentioned in other modes. For example, the user can combine text input and mouse pointing following Bolt's (1980) Put-That-There system (e.g." Is using thin cable possible in <mouse select> this shaft?") and the system can combine graphical output with text (e.g. "What is the type of <system highlight cable> this cable?").

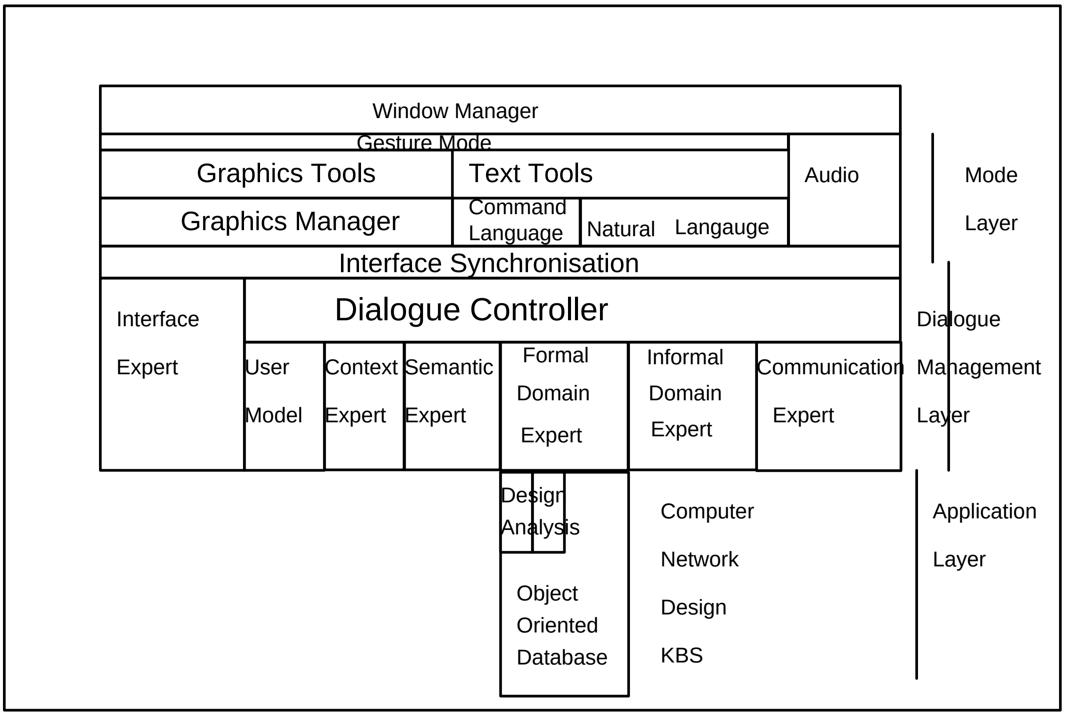

Figure 3: Architecture of the MMI^2 Multimodal Man-Machine Interface for Knowledge Based Systems.

MMI^2 does not include a broad knowledge base of common-sense knowledge (e.g. Guha & Lenet, 1990) but it must include more than just the limited domain knowledge for the demonstrator application for designing computer networks. Two other domains of knowledge are represented: the domain of the user, and the domain if the interface itself.

The user model contains a model of the beliefs of the user (Chappel et al, 1992). It monitors all messages passing between the mode and dialogue management layers in CMR and extracts from them beliefs which the user holds (both correctly and incorrectly with respect to the knowledge stored in the KBS in MMI^2 which are assumed to be correct), and the intentions of the user. This user model then acts as a server to other parts of the system which require knowledge of the user, such as the graphics manager for planning effective graphics communication, the natural language generators for generating text, and the communications and informal domain experts for planning multi-modal fission. The user model is also available for the evaluation or interpretation of predicates in CMR about the user (e.g. to answer questions such has "Who am I?").

Figure 4: An example screen image from the second MMI2 demonstrator for monitoring computer network performance showing different interaction modes: a map overlaid with the physical structure of the network; natural language input mode, the logical structure of the computer network.

The interface expert contains information about the interface itself. This is available to answer questions about the interface and its capabilities, but also for the evaluation of predicates in the CMR about the interface. For example, if the user commands the system to "draw a bar chart of the cost of computers on the network" then concepts such as BarChart are not network design concepts, but interface concepts; so that their evaluation is against the domain of the interface rather than network design.

The third domain is obviously the domain of the application itself, containing knowledge of computer network design. This application is accessed through the formal domain expert which provides a functional interface consistent with the user model and interface expert. The application itself consists of an object oriented database which stores and object and instance hierarchy for the computer network domain ontology. In addition there are two sets of domain heuristics, for design, and for the analysis of a design. These were developed using task models for Hierarchical Design and Heuristic Classification operating on the domain objects in the object-oriented database. Above the Heuristic Level of the application itself, the Formal Domain Expert represents the CMR predicates corresponding to application functions.

The functional interface provided to the knowledge bases in the Formal Domain Expert, User Model and Interface Expert includes three operations: Assert, Retract and Goal, to update the knowledge bases, retract information from them, and ask questions of them. This is similar to the level of operation provided in other knowledge based systems (e.g. Guha & Lenet, 1990).

The use of three domains of knowledge through a common functional interface allows the dialogue controller to be efficiently implemented and the system to be extensible through this common interface. The different domains allow users freedom to act and ask about more than just the core domain itself and provide some impression of cooperativity. Unfortunately, the addition of these two domains alone do not achieve the structure of a co-operative human-human conversation (e.g. Sacks et al, 1974). To do this requires a representation of the application task strategy and knowledge of communication argumentation to present the information to the user co-operatively. An example will illustrate this point more clearly. When a user wishes to design a network they must state some essential requirements such as a description of the building, the number of machines required, and the cost of installation. There are also many optional requirements such as demands to promote extensibility, and constraints on some environments being hazardous to network performance (e.g. X-ray exposure). If the user fails to state all of the essential requirements, and asks the system to perform a design, then it will merely fail to produce a design. It is necessary to provide an explicit high level representation of the task model to ensure that the user is prompted for requirements, and that the requirements stated are not contradictory. These task plans are represented in the Informal Domain Expert. Before Goal predicates with free variables are interpreted against the Formal Domain Expert they are passed to the Informal Domain Expert for informal evaluation that pre-conditions on the task stage have been met. If they have not, then messages are passed back to the Dialogue Controller.

In order for the Dialogue Controller to express these or other statements to the user, they must be structured into complex messages to present the argument that is required in an effective way. To do this MMI^2 includes a communication planning module which turns sequences of functional interface level formulae into large CMR structures which can be passed to the modes by planning the argumentation structure of the message using knowledge of communication itself (Cohen, 1992 provides an introduction to rules for selecting graphical and textual modes). The communication planner internally uses further abstract levels of representation of interactions in the form of communication forces (following Maybury, 1991).

This description of MMI2 has shown that co-operative multimodal dialogue can be demonstrated in an architecture guided by five principles:

1) The use of as single common meaning representation formalism, common to all modes, which is used as a vehicle for internal communication of the semantic content of interactions inside the interface and also as a support for semantic and pragmatic reasoning.

2) Mode integration should mainly be achieved by an integrated management of a single generalised discourse context.

3) There are different model theories for the evaluation of symbols in the meaning representation formalism for the application, the interface domain, and the user domain.

4) The effect of formally evaluating communication actions against a domain can cause side effects in other domains.

5) The use of pragmatic task plans in the informal domain expert and expertise about communication planning in the communication expert, graphics manger and natural language modes to produce high quality, effectively planned communications through informal processing.

Within the MSM framework, MMI2 is clearly not only a multimedia, but obviously a multimodal system in that it: a) supports both the fusion of user input from different modes (using the dialogue controller and context expert) and the fission of its output between the most effective an efficient modes (using the communication planner); b) it provides several communication channels which operate, c) between both the user and the system, and the system and the user; d) it uses raw data representation, but also a common meaning representation and communication forces in communication planning; e) it uses the context of the task, user, and dialogue to tailor output and interpret input; f) it allows various levels of parallelism including physical actions, task and task cluster since multiply nested dialogues are supported.

Complete multimodal systems such as MMI^2 are currently only research demonstrators which can produce potent illustrations of multimodal interaction, but are not even robust enough for real user evaluation. However, many components shown in this system are being brought to the marketplace where there is seen to be a need for them. Gesture interfaces of the form used in MMI^2 have now been incorporated in personal digital assistants (PDA) for recognising symbols, even if they are not practically sophisticated to interpret cursive writing sufficiently reliably yet. The freehand input of building drawings and their automatic interpretation into objects used in MMI^2 has also been incorporated in many PDAs. Companies such as Apple and Psion are developing speech input command systems which will fuse their input with that from gesture mode using principles developed in MMI^2. Many spreadsheet developers are including business graphics creation rule sets to automatically generate bar charts, pie charts and graphs from spreadsheet data using rules similar to those used in the MMI^2 graphics manager. Constrained natural language query systems for databases incorporate technologies which are a subset of those used in the MMI^2 natural language input mode, and many of these are starting to incorporate graphical interface tools to help complement natural language, combining modes again using principles seen in MMI^2.

Despite the incorporation of parts of the MMI^2 functionality in commercial development projects, there are many problems still to be overcome with the complete system. Further research is required in focused projects to refine the output generation rule sets, the context sensitive reference resolution rules, to improve the efficiency whilst maintaining the expressiveness of the meaning representation language and mainly to improve the development method for capturing the requirements for co-operative dialogue. It is not practical to move from current application architectures to logic based dialogue centred systems such as MMI^2 in one step. It is necessary to isolate out those aspects which can be combined into more conventional designs to add functionality to them in order to dynamically create presentations which are tailored to the user, task and dialogue context as those in MMI^2 are.

MIPS

As stated above, current hypertext systems are closed individual products where the reader can choose routes through the web, but the content has been chosen, and the link design completed by the author. A step towards opening up such systems is provided by the Microcosm system (Davis et al, 1993; Hall, this volume) which separates the link structure from the data assets presented in the hypertext thereby allowing users to link pre-existing documents, images and other media items to a web. Since media items are independent they can also be stored in the formats of common presentation tools, providing users with a more consistent system image between the hypertext and other tools on a system. This changes the hyperdocument from being purely an artifact created by an author to one which can be part of an open information system where assets can be re-used in many documents (easing potential copyright problems too).

The next stage in opening up hypertext documents is to represent the link structures themselves in an interchangeable language. This would allow a set of data assets and the link structure to be portable across different presentation and link authoring platforms. This standardisation of open link representation is available in the ISO HyTime standard for hypermedia time based data.

These two advances of opening hypertext by separating assets from links and then using standardised representations for both assets and link webs support the portability of hypertext. The next stage in opening hypertext and moving away from the publishing model of authored and read product is to introduce some connectivity into the presentation system to allow nodes, links and screens to be produced dynamically by the reader. The product GainMomentum from Gain Technology is one of the first hypermedia systems to include database querying and the production of screens from the returned data as an integral component. This allows authors to include queries to on-line databases in the web so that a node can be populated at query time. This can be very useful when tables such as railway timetables need to be included in screens. The query can be issued by the system when the reader asks to move to a node on the topic, the data will be returned from an on-line database and included into the template for the node and screen for presentation to the user. This allows the tables presented to the user to be up to date with respect to the current timetable database rather than having the author produce timetables at authoring time which may be out of date by the time of reading.

The next step in the development is to allow not only pre-written screens to be populated by the returned data, but also to have the node, links from it, and the screen designed on-line in response to the query and the returned data. This allows the presentation to be more flexible and to account for variation in the data which is available at run time but which may not have been entirely considered by the author. The MIPS system takes this step and introduces a loose template system with a KBS presentation design system to construct nodes, links and screens from returned data. Following these stages of development of open hypermedia, MIPS uses pre-existing asset formats and existing presentation tools for those as Microcosm did, it is based upon the HyTime standard rather than a proprietary link representation system to further promote portability, it includes not only SQL queries as GainMomentum does, but also document retrieval interfaces so that assets of all presentable media can be retrieved.

Any hypermedia product at present must be able to support the conventional publishing life cycle of authoring, distribution and reading without adding intelligence in itself. Therefore the MIPS system is must be capable of this too. In order to present pre-written hypermedia documents without any queries, presentation tools, a presentation manager to control the browsing of the web structure and the delivery of data assets to those tools, and a storage mechanism for the hypermedia web are required. The shaded area in Figure 5, shows the architecture required for these functions: presentation tools, a presentation manager, and a HyTime Engine to provide fast access to the object oriented database containing the web link structure.

In order to connect the hypermedia tool to databases so that pre-written queries in the web can be used to existing heterogeneous databases, a selection and retrieval tool to select the appropriate database as the target for a query, and to format the query for that database is required. Once the query is formatted it must be dispatched to the remote or local database through a communications module, and the returned data must be passed back to the selection and retrieval tool. This must be incorporated into the HyTime web representation by a Web Builder, so that the node can be presented. Again, in Figure 5, the modules to provide this functionality are also shown: Selection and Retrieval Tool, Communications Module and Web Builder.

The third step of adding dynamic construction of screens in response to user queries at run time rather than pre-written queries produced by an author requires a Query Tool for the user to express queries in (also shown in Figure 5). In order to perform the selection of the appropriate database from the heterogeneous set available the system must know what databases are available, what information they represent, what format queries and returned data use, and other information about cost, access time and other non-functional requirements in order to optimise the query. This information is stored in the Knowledge Based System which supports the Query Tool and Selection and Retrieval Tool in the construction of queries. Similarly, the returned data must be constructed into a node, which must be linked to other nodes, and then presented on the screen using the most effective and efficient presentation mechanisms available for that class of data. Again, information design expertise to support this task is stored in the knowledge base system which supports the web builder in constructing the node, and the Presentation Manager in selecting the most efficient presentation tools to render a presentation mechanism.

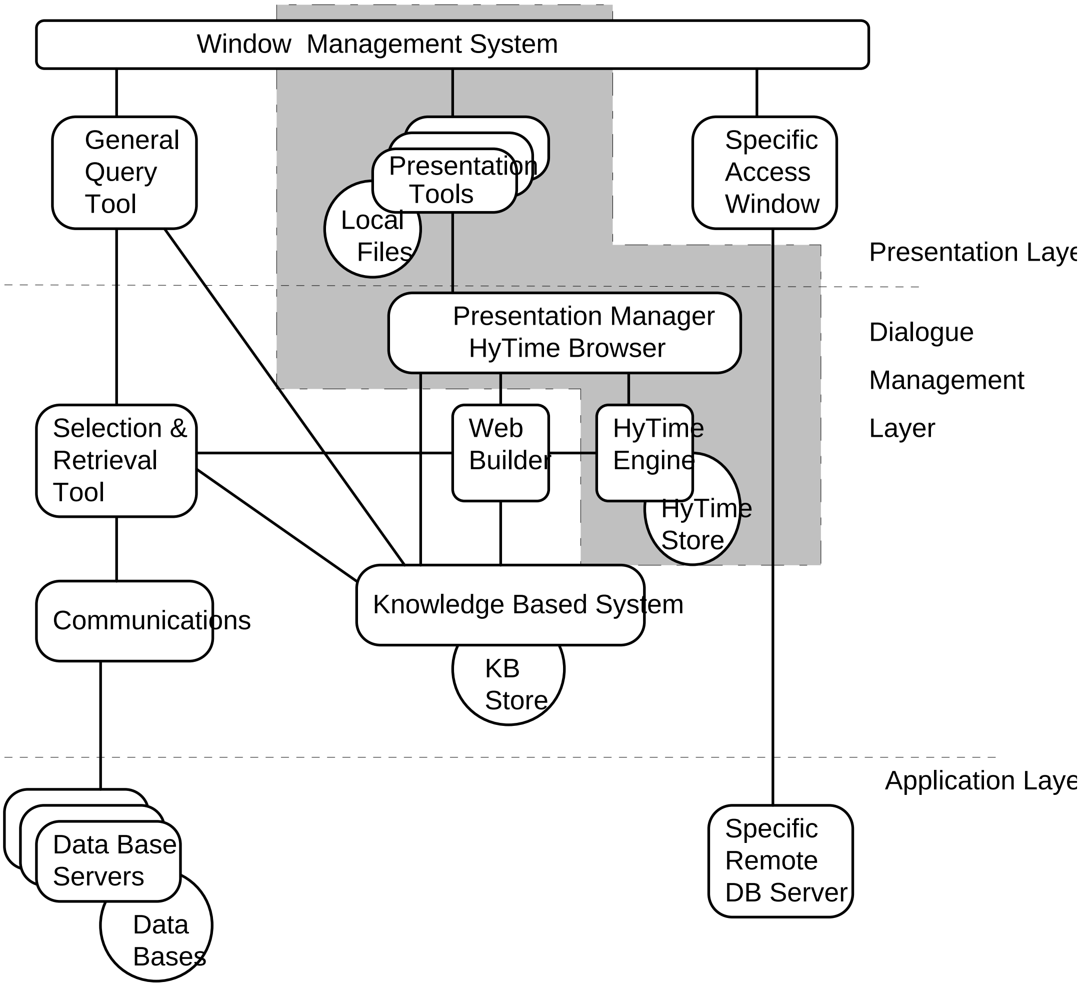

Figure 5: Architecture of the MIPS multimedia presentation system.

As with the MMI^2 system, the architecture is clearly divided into three layers responsible for the presentation of information, the dialogue management and the application itself. The application in this view of MIPS is the local and remote databases. This clear division is a little confused by the existence of two permanent stores in the dialogue management layer: the HyTime store, and the Knowledge Base; and one in the presentation layer: the files of local data assets for presentation. Given current hardware, the size of video and sound data files, and the speed of transferring them, it is necessary to keep large data assets as close as possible to the presentation tools to provide the speed and quality of presentation demanded by users. MIPS is designed to be run as a client server system with several client machines running presentation tools, with one server responsible for the database access. Therefore to keep data assets close to the presentation tools they are stored on the client machine on which they run, although they can be regarded as data assets cached from those held in databases. The dominant dialogue control mechanism is the web structure itself which restricts the links available to users within the web. Although the hyperdocument may be considered as the application, as an application divided between assets and web, the sole role of the web is to control dialogue. This is therefore represented within the dialogue management layer. The KBS store contains knowledge used by the KBS to select databases and design nodes and presentations. These functions are comparable with the informal domain expertise and communication planning functions in the MMI^2 system and for the purpose of the main application function are also dialogue management functions so their placement in the dialogue management layer is appropriate.

One of the objectives of the MIPS system is to improve the portability and interoperability of hypermedia systems. Portability is addressed by using existing data formats and presentation tools, and by using a standard representation for the web. To promote interoperability knowledge about the applications with which it must operate has been included in the dialogue management layer, within the KBS: about databases to which it is connected, and about presentation tools available at the local site. Also knowledge has been included about the presentation of data assets, and the construction of nodes, links and presentations, so that data can be integrated into the hyperdocument web. There are several discrete bodies of knowledge that have been included within the KBS, but unlike the MMI2 architecture where the knowledge is distributed around the system most of it is a KBS written in Prolog, in MIPS all this knowledge is placed in a single KBS separated from the rest of the architecture. This allows the KBS to act as a server to the rest of the system and for a core presentation only system to operate without it. However, this does not constrain the complexity of the KBS itself which is now internally modularised.

Figure 6: Screen design for a hypermedia tourism application for the Barcelona 1992 Olympics using the MIPS presentation system. The main image is a short video of site.

There are four domains of knowledge which the KBS must know about to perform its functions. Firstly, it must contain knowledge about database access in general and the optimisation of queries. Secondly it must understand the principles of information design in order to construct nodes, links and screens. Thirdly it must understand the language of the application domain since actual queries to databases in an application will use this information, and the actual data to be presented in an application will be in terms of this information. Fourthly, it must have knowledge of the actual presentation tools present at a site and the rules for choosing these for a presentation mechanism. These four domains of knowledge interact so that there is general knowledge about database access and information design, then layered upon this there is knowledge specific to the application domain. The knowledge which was spread through the communication planner, graphics manager and interface expert of MMI2, here all resides in information design and presentation knowledge. What were the domain experts in MMI2 here become knowledge about the domain of database access, and the application domain. In order for the KBS to perform its functions, other knowledge which was present in MMI2 must also be included here. In order to design information it is again necessary to represent the dialogue context, and to classify the user in a user model. Similarly, to permit any portability of the

KBS it is necessary to divide four domains of knowledge between task models which describe the functions to be performed in an application domain platform, and then layer application domain knowledge on top of this in a conventional KBS domain model.

These various layers of knowledge required to automate presentation design must be acquired, and cannot all come from an application builder or author who will not be a knowledge engineer. General knowledge about database access and information design exist within the basic MIPS KBS. There is also a core ontology of terms defined which are used in the MIPS DTD for HyTime documents produced for MIPS. On top of this it is necessary for the application builder to extend the ontology for the application domain through a simple tool designed for this purpose. The domain specific rules for database selection and information design can also be entered by the application builder (author) since only a very limited set of these beyond the ontology are required (e.g. constraints on some transitive inferences such as 'travel from A to B, via how many intermediate places'). It is necessary to extend the directory of remote databases available for an application, and indicate the schema of these using terms in the ontology, but this is a form completion exercise and does not require knowledge acquisition. The application builder will also specify the simple user modelling structure by choosing categories of features which can be associated with different user groups in the ontology. The application builder will therefore be required to enter some information into the KBS, but form interfaces are provided to limit the complexity of this operation. These must be evaluated before the system could be viable, and they will undoubtedly have to be changed, but the present system is intended to address this potential problem of overloading the application builder. Once an application has been written in a domain (e.g. Greek tourism) very small additions would be required for further applications. Therefore the knowledge can be seen as a general layer, one for the application domain and a third for each specific application.

The application builder would obviously have to author the HyTime web for an application as well as entering the KBS knowledge. It is probable that some data assets would also have to be created for an application, although these could already exist or be the responsibility of database providers.

Once the application has been created it must be installed at each site. Another layer of customisation is required here to state which actual databases are connected to it, which actual presentation tools are present at the site and which actual user groups are potential users. The preferences of the user groups can also be tailored to the site. this provides a fourth layer of customisation to the user site. The fifth layer of customisation available is to each independent user who can customise their own user model to include their own preferences for database access variables (cost, time etc.), presentation tools to be used, language etc. These can be defaults for users at a site, or can be set explicitly by users willing to devote the effort to this task. The sixth level of customisation available is to an individual session's dialogue context. This is automatically performed by the KBS which keeps track of the context throughout use and automatically updates the user model of a user on the basis of usage.

These six layers of the customisation process provide the knowledge on which the system can base its retrieval and information design judgements and tailor them to individual users so that they appear intelligent. To perform these customisation steps three different groups of users have been distinguished: application builders who develop the application, site managers who configure the application to a site, and end users who configure the system to their own preferences. Other groups of individuals could also be involved such as specialist asset generation teams (including film directors, graphic artists etc..), remote site database managers, or even domain modellers.

In this application both the problems of heterogeneous database access and those of information design must be addressed. A major worry is the automation of these two areas is the need for general knowledge and ontologies. Stated boldly, the generation of general knowledge systems is a long term goals of the artificial intelligence community and should not be regarded as a solved problem. Within the limited application supported by the MIPS system, and with the deliberate exclusion of any natural language input it is hoped that these issues have been sufficiently constrained so as not to be onerous to application builders. If this is proven not to be so in evaluation, the approach should not be rejected immediately. Artificial intelligence workers such as Guha and Lenat (1990) are developing exactly the form of general knowledge base which could be used to support heterogeneous database access (Lenat & Guha, 1991) and may provide a general basis for the link creation part of the information design process in MIPS applications.

MIPS supports the retrieval of any information stored in data sources. One practical limitation on this is the communication of large media (i.e. video, sound) over computer networks. If this is not practical in real time then the system is only useful for retrieving less bulky media. This is an immediate problem, but even now there are local uses for this technology if not over wide area networks. A second problem with the retrieval of such media is the indexing of it in databases. There are currently no clear standards for the storage of multimedia data, nor for querying it. Querying may seem a strange point to raise, but if a video display is being automatically composed, it could also be automatically edited. That is, a ten minute video of a scene could have a shot extracted from it, or a series of shots which could be put together into a three minute sequence. If the video were indexed using techniques such as that proposed by Burrill et al (in press), then it could have an index attached to each scene and shot not only with the identity of characters, locations and actions in it, but also video attributes such as whether it was an establishing shot, close-up, pan etc. Such indexing on semantic content and video attributes would support the retrieval and composition of sequences according to rules of video direction. This is one example of the forms of presentation design which could be developed for multimedia systems using a MIPS architecture as the starting point; although it has not been developed yet.

The MIPS system is a demonstrator which is considerably simpler than MMI2. Within the MSM framework of Coutaz et al. described above, it supports only one channel from the user to the system since all user input is through keyboard or mouse selection, but it uses visual and auditory channels for presenting information to the user. It does not support the fusion of user input, but since it can produce several media in reply to a query it could be said to support fission. It has been designed as a multimedia rather than a multimodal system which does not support advanced user input dialogue so this judgement is not surprising. Similarly, parallelism is only supported in system output at the physical level and could only be said to be supported in user input at the task level if queries are pending the return of data while a user continues to browse the hyperdocument. The abstraction of the dialogue supported is more problematic. Outside the KBS there is no abstraction of navigation commands to the web, and little abstraction in data source queries. Within the KBS there is some abstraction of the query to the user's task level but this is mainly to support the use of contexts. The MIPS KBS is designed to accommodate the context of the task, user and dialogue in designing presentations for returned data and linking these into the web. It is this use of contexts in presentation generation which is the contribution of any intelligence in this system. The system has not yet been implemented to a stage which will support evaluation so it cannot be judged whether this is enough. However, an evaluation will take place in Greece of an application in the tourist domain in order to answer this question.

Conclusion

Current multimedia systems present text, images, sounds and video as static objects. The organisation of these objects is determined by the author, and the reader has a comparatively passive role. In order to involve multimedia in more applications it must become more active with the representations used must include meaning which can be manipulated in more complex dialogue structures. The technology currently used to support multimedia presentation does not incorporate any analysis of the signals, any discussion of human-computer communication modalities inherently involves, at some level, the machine's determination of the content of messages, and its need to communicate the content of its own messages. Multimodal systems use different media for input and output, and rely on abstract representations of the information in order to control the dialogue and application. An example multimodal system was described - MMI^2. Such systems are far from the current market but they provide the basis for determining the theory of multimodal communication which is required to be stated in detail if multimedia presentations are to be automatically created.

A second less advanced multimedia system (MIPS) has been described which is an advance on current hypertext, including some intelligence to support the dynamic creation of multimedia documents from retrieved data. This illustrates a second route towards the introduction if intelligent multimedia through open systems which interact with existing data sources. Although this system does not include the advanced dialogue of a multimodal system, its use of context information to dynamically create presentations may incorporate the most robust aspects of the more exotic system in a way which can be included in commercially supported products.

It is arguable that neither of these systems portrays intelligence as the title suggests. This conclusion is either drawn because of a conviction that it is a misnomer to term any system intelligent or because there are few examples presented that show the systems in operation. In the first is true, the term is used suggestively rather than with any psychological conviction. If the second is the case, on several occasions when MMI^2 has been demonstrated to computer professionals they have refused to believe that it was interpreting the user's input or generating its own output. They claimed that there was either somebody behind the curtain, or the demonstration was so well prepared that the system could not stray outside it. It is this impression of disbelief in those who see the systems which leads to their being called intelligent. Their developers do not make any greater claim.

This chapter has not described various rule sets for dialogue context, user or task modelling, for generating output in different modes nor for selecting media, but references have been provided to sources of this information for those who wish to acquire it. These rule sets are still active areas of research and some are being incorporated into existing major software products to provide them an edge in the current market.

Acknowledgements

The research reported in this paper was partly funded by CEC Esprit II grant 2474 to the MMI^2 project and partly by CEC Esprit III grant 6542 to the MIPS project. Organisations involved in the MMI^2 project are BIM (Belgium), ISS (Spain), ADR/CRISS (France), INRIA (France), EMSE (France), SERC RAL (UK), University of Leeds (UK). Organisations involved in the MIPS project are.Longman Cartermill (UK), SERC RAL (UK), STI (Spain), SEMA Group Benelux (Belgium), Herriot-Watt University (UK), Trinity College Dublin (Eire), DTI (Denmark).

References

Binot, J-L., Falzon, P., Perez, R., Peroche, B., Sheehy, N., Rouault, J. and Wilson, M.D. (1990). Architecture of a multimodal dialogue interface for knowledge-based systems. Proceedings of the Esprit '90 Conference, pp 412-433. Kluwer Academic Publishers: Dordrecht. Also published on CD-ROM by the CEC: Brussels.

Bolt, R.A. (1980) "Put-that-there": voice and gesture at the graphics interface. Computer Graphics, 14(3), 262-270.

Bush, V. (1945) As We may Think. Atlantic Monthly, July, 101-108.

Burrill, V., Kirste, T., Weiss, J. (in press) Time-varying sensitive regions in dynamic multimedia objects: A pragmatic approach to content-based retrieval from video. Software and Information Technology - special issue on Multimedia.

Chappel, H., Wilson, M. and Cahour, B.(1992) Engineering User Models to Enhance Multi-Modal Dialogue. In J.A. Larson and C.A. Unger (Eds.) Engineering For Human-Computer Interaction. Elsevier Science Publishers B.V. (North-Holland): Amsterdam, pp 297-315.

Chappel, H. and Wilson, M.D. (1993) Knowledge-Based Design of Graphical Responses. In Proceedings of the ACM International Workshop on Intelligent User Interfaces, pp 29-36. ACM Press: New York.

Cohen, P.R. (1992) The Role of natural language in a multimodal interface. In the Proceedings of the Fifth Annual Symposium on User Interface Software and Technology, 143-149, ACM: New York.

Cotton, R. and Oliver, R. (1993) Understanding HyperMedia, Phaidon Press: London.

Coutaz, J., Nigay, L. & Salber, D. (1993) Taxonomic Issues for multimodal and multimedia interactive systems. In Noëlle Carbonnell (Ed.) Proceedings of the ERCIM Workshop on Multimodal Human-Computer Interaction. INRIA: Nancy.

Crane, H.D. & Rtischev, D. (1993) Pen and Voice Unite. Byte, 18(11), 99-102.

Davis, H., Hall, W., Pickering, A., and Wilkins, R. (1993) Microcosm: An open hypermedia system. Proceedings of INTERCHI '93 Conference on Human Factors in Computing Systems, ACM: New York.

Feiner, S. and McKeown, K. (1991) Automating the Generation of Coordinated Multimedia Explanations. *IEEE Computer* 24(10); 33-41.

Guha, R.V. and Lenat, D.B.(1990) Cyc: a midterm report, *AI Magazine*. 11 (3), 32-59.

Haynes, S.L. (1991) Intellectual Proprty and Licensing Concerns. In E Berk and J Devlin (Eds.) Hypertext/ Hypermedia Handbook, McGraw-Hill: New York, NY, 227-241.

ISO (1992) ISO/IEC JTC1/Sc18/WG8, Information Technology, Hypermedia/Time-based Structuring Language (HyTime). ISO/IEC D18 10744.1.1.

Lenat, D.B. & Guha, R.V. (1991) Ideas for applying Cyc, Tech Report. ACT-CYC-407-91, MCC: Austin TX.

Lyons, P. (1991) Copyright considerations of Hypertext producers: Imaging and Document Conversion.In E Berk and J Devlin (Eds.) Hypertext/ Hypermedia Handbook, McGraw-Hill: New York, NY, 259-267.

Mackinlay, J. (1986) Automating the Design of Graphical Presentations of Relational Information, ACM Transactions on Graphics, 5(2), 110-141.

Maybury, M.T. (1991) Planning Multimedia Explanations Using Communicative Acts. In Proceedings of the Ninth national Conference on Artificial Intelligence, AAAI-91. Morgan Kaufman: Los Altos CA.

McIntosh, S.I (1991) Intellectual Property Issues in Multimedia Production. In E Berk and J Devlin (Eds.) Hypertext/ Hypermedia Handbook, McGraw-Hill: New York, NY, 243-251.

Nelson, T.H. (1988) Hyperdocuments and how to create them. Prentice Hall: New Jersey.

Newcombe, S.R. et al, (1991) The HyTime, *Communications of the ACM*, Vol 34 (11).

Nickerson, R.S. (1977) On conversational Interaction with Computers, in User-Oriented Design of Interactive Graphics Systems. New York: ACM.

Sacks, H., Schegloff, E. & Jefferson, G. (1974) A Simple Systematics for the Organisation of Turn-taking in Conversation. Language, 50, 696-735.

Tansley, D.S.W. & Hayball, C.C.(1993) *Knowledge Based Systems Analysis and Design: A KADS Developer's Handbook*. Prentice-Hall: London.

Templeton, A. (1993) Strategies for Business, In *European Multimedia Yearbook 93*, Interactive Media Publications: London.

Brief Biography of the Author

Michael Wilson is a Chartered Psychologist holding BSc and PhD degrees in Experimental Psychology. Since 1983 he has undertaken research into Human Computer Interaction at the MRC Applied Psychology Unit, Cambridge and Knowledge Engineering at the SERC Rutherford Appleton Laboratory where he is currently head of the Intelligent User Systems Section. At present he is the

RAL team leader in the Esprit projects MMI^2 and MIPS, and acts as a monitor of research for the SERC. He has published over fifty book chapters and journal articles on task analysis, modelling human-computer interaction, knowledge acquisition, and multimodal and multimedia user interface design.