2

Give us the tools: a personal view of multi-modal computer-human dialogue

David R. Hill
Department of Computer Science
and Department of Psychology
The University of Calgary
CALGARY, Alberta
Canada T2N 1N4
email: hill@cpsc.ucalgary.ca

1. Introduction: information, knowledge technology, and interaction

1.1 Brief origins and context

Over the past three decades, computers have become to our brains, what such inventions as mechanical conveyances are to our legs, and optical instruments to our eyes. Naturalness, elegance, functionality, reliability, and ease of use are essential to achieving the basic usability goals of these neuro-sensory-motor extensions—namely to make their power easily available to whoever needs to use it for whatever job is to be done, in a non-intrusive manner.

When automobiles and telescopes were first invented, the problems of making their powers available were largely restricted to the physical domain. The term *telescope* has become synonymous with reducing or making smaller, as well as naming the instrument, because the most important problem with early telescopes was to make them small enough to be carried conveniently.

The problem of making computer power available goes well beyond the physical. Even now, when computers are small enough to fit in the pocket or on the wrist, accessibility, through some kind of part-real, part-virtual physical interface (Computer-Human Interaction—CHI), is a serious challenge, and is still the focus of intensive research efforts.

Debates such as those between supporters of *Apple*, *NeXT* and *Microsoft* interface implementations, as well as the gap between these commercial interfaces and the technology that can be seen at professional conferences, reminds one of the early days of motoring when most automobiles had pedals intended to control the vehicle, but the pedals were differently arranged and did different things depending on the manufacturer. Not all vehicles had steering wheels. Such variety is almost certainly essential during experimental development, but provides a barrier to convenient use. Of course, early standardisation can have the same end result and, one hopes that we shall avoid "QWERTY keyboard-type" solutions to modern CHI problems.

The term *multimodal* appears increasingly often in the context of the computer-human interface. There is a plethora of words to describe different aspects of the emerging new interface technology. They are often used interchangeably and inconsistently. At this workshop, terminology and methodology have sparked lively debate and a realisation that work on a common language for our research is important. Multimodal seems to mean interfacing using more than one sensory-motor modality (touch, sight, etc.); multimedia is more concerned with the integration of different media types; but terms like "mode", "channel", "medium" and "protocol" are not used consistently and some (the present author included) are adamantly opposed to the idea of using words like "protocol", with their connotations of rigidly specified, non-redundant forms of communication, for any normal human communication, whether with another person or a machine.   Thomas (1978) and Thomas and Carroll (1981) have eloquently explained why communication should be regarded as a "Design-Interpret" activity rather than an "Encode-Decode" activity, and thus thoroughly undermined any idea of using layered protocols as a model for any communication involving humans. At the same time, one may be able to draw analogies between human communication and formal communication theory, provided the limitations of the analogies are kept firmly in mind. Human communication is concerned with purpose, and involves context (pragmatics). Shannon clearly excluded any level above the syntactic from the domain of his theory (Shannon 1948), and little progress has been made in attempts to extend it.

1.2 AI and CHI

Computer interfaces have always been multi-modal, with both formal and informal channels. We have to ask: "Why the new emphasis?". I suggest that it has to do partly with crossing a threshold of complexity in much the same way that the older discipline of Human Factors (Ergonomics) really took off—quite literally—when complex equipment such as the modern aeroplane made its appearance around the time of the second world war (say 1941), taxing human sensory-motor skills to the point where new knowledge and techniques were required for successful design and operation. This crisis coincided with an important breakthrough, that of Electronic Device Technology—the application of vacuum tubes to computing and other digital information processing tasks, and the

conception of the General Purpose Digital Computer.   This early work on the scientific study of the relations between people and increasingly complex machines may properly be called *Phase 1* of the emerging technology, as the twin roots of modern CHI were formed and became intertwined over the next two and a half decades—computers and human factors.

Phase 1 started with largely independent developments in these two areas, resulting in computers that were accessible only to highly trained operators and programmers. However, the new means of information processing led to two hot new areas of research. One was concerned with automating "thinking processes" (i.e. creating forms of "artificial intelligence" or AI). The first landmark international conference in the area took place at the National Physical Laboratory, in England, in November 1958 (NPL 1959), concerned with topics like neural networks for pattern recognition and learning, automation of programming, modelling biological systems, and language translation. The other, slower to start, was concerned to develop better ways of interacting with computers. This was not initially considered to be part of human factors, which was by then established as a branch of psychology.   The invention of time shared computers, leading to multi-access interactive computing by individuals, and the development of cheaper, more powerful hardware, leading to laboratory computers like the PDP-8, that were used by people whose expertise was concentrated in professional areas other than computing, led to another crisis in human-machine relations.

Nilo Lindgren, writing at the end of Phase 1, expressed the crisis succinctly:

> *Now the emergence of the computer sciences, in which the human*
> *characteristics must be matched with the machine at the intellectual*
> *and deeper neural levels, threatens to place new burdens on what*
> *human factors means, as a name. (Lindgren 1966, p 139).*

It was no longer enough to make sure the keys were appropriately designed and the displays legible. The information had to be presented in a human-compatible manner, and the input adapted to the way humans understood the material. As Lindgren wrote, *Sketchpad* had just been developed by Ivan Sutherland at *Lincoln Lab* and *MIT*, Doug Englebart had invented the mouse and gestural input at *Stanford Research Institute* (*SRI*) in Palo Alto, and Alan Kay was developing "*Dynabook*". Even today, few graphics programs embody all the innovations from Sutherland's program, we have barely scratched the surface of the possibilities introduced by gestural input, and *Dynabook* is still a glimmer in its inventor's eye—despite the fact that most of the ideas have been imported into other systems and despite the excellent laptops and palmtops now available.

Lindgren's observation marked the start of Phase 2 and also the start of the on-again, off-again relationship between AI and CHI. Many of the ways in which researchers wished to provide innovative computer-human interfaces, or to extend the autonomy of computers to make them more useful/usable, required solution of AI problems, especially the "AI-hard" problem of language-understanding/world-knowledge-representation

Thus, human-machine relationships could no longer be dealt with entirely in terms of physical, anthropometric and psychophysical terms, but would increasingly have to be dealt with using the new tools of cognitive psychology and artificial intelligence.   This led directly to the development of computer-human interaction (CHI) as a separate discipline from more traditional human factors, and to the revolution in ease of use of information processing systems that took place during the 70s and 80s, as expert systems, and "intelligence-in-the-interface" to support the user became more common. However, AI and CHI researchers split during Phase 2, and pursued their own agendas and methodologies, with the artificial intellingentsia worrying about theorem proving, learning, language understanding and the like, while the "CHIborgs" studied the problems of computer displays, command naming systems, programming, and user interface prototyping and management.

It wasn't until 1982 that the first major conference devoted to human performance issues in computer systems design, development and use was held at Gaithersburg, Maryland, sponsored by the National Bureau of Standards, and the ACM (NBC/ACM 1982), and the topics were still heavily oriented towards the psychology of interaction.

By the end of the 80s, many of the ideas developed in the research labs such places as Xerox PARC, Lincoln Labs and the *Stanford Research Institute* (now *SRI International*) were incorporated into off-the-shelf products available to consumers. Computer power had increased by five orders of magnitude in terms of speed and complexity, whilst decreasing in size by three or four orders of magnitude since the beginning of Phase 1, and what once was beyond the reach of the most powerful governments, or later took a research supercomputer, was now conceivable on a personal workstation or laptop.

As another 25 years passed and we entered the 90s, a new crisis loomed. This crisis was caused by the impossibility of managing, within the confines of traditional CHI, the new complexity.

People wish to connect their computers to networks and enter virtual realities of one form or another, using communications and media to eliminate restrictions of time, of space, or of intellectual and physical ability. In the 90s we see Virtual Reality (VR) of the most complex and sophisticated kind as a topic of research in its own right, but VR is a continuum. Even the simplest computer-human interface that is not strictly physical is a minimal form of virtual reality. Perhaps one of the less considered virtual realities is that provided by the cockpit of an aeroplane under instrument flight conditions, especially with the advent of aircraft like the A300 Airbus in which the controls and displays all attach to a computer, which does the actual flying. At least with aircraft, there is a well-defined paradigm, plus 50 years experience, for the creation and management of the partial artificial reality that is created. But Virtual Reality is fast becoming the paradigm for CHI in general, and it seems no accident that VR featured prominently at SIGCHI 90 in Seattle, as the decade opened and Phase 3 got under way. But there are more Virtual Realities than those that require a headset and data glove, and the reality may be truly artificial, in

the sense that it is an unfamiliar or managed reality [1]. Managing the resources needed to create such systems is proving to be a very serious problem. This is the real difficulty with multimedia. Multimedia also requires that we gain insight into how we think, understand and communicate as humans, not necessarily in order to imitate, but certainly in order to accommodate. In Phases 1 and 2 we were largely able to ignore these fundamental problems which are the hardest problems in AI. It is these problems of Phase 3 that our workshop seeks to address.

Users today are increasingly sophisticated because they are more UI-educated, more diverse, less tolerant and are driven by companies eager to apply computers to an ever extending range of tasks and activities, in order to generate new products. To meet these needs, as well as to meet the basic requirements of excellent user interface design more completely, increasing amounts of intelligence and autonomous behaviour must be built into the systems. Intelligence in the interface manages resources, provides new channels for communication, co-ordinates different media, different modes and different physical devices. Whichever way you cut it, whether in creating new tools such as those needed for speech input and output, or in managing partnership dialogues, or in providing reasoned access to databases, or in knowing how the signals in different computer input modalities relate to each other (to mention only a few tasks that spring to mind), CHI has to renew the marriage with AI that was largely sundered in Phase 2. This will be the pressing task of Phase 3.

1.3 BRETAM: the substructure of cultural advance

In a number of papers Gaines has proposed and explored a model for cultural/technological advance (for example, Gaines & Shaw 1986; Gaines 1990). Figure 1 shows the basic BRETAM model which proposes that human technological progress proceeds rather like a series of thundershowers of creativity, in which each thundershower sows the seeds for a successor. But each thundershower itself is also structured.

At our workshop, Martin Taylor remarked at breakfast one morning that we were talking about the same things at Maratea as we had talked about at the original workshop in Venaco five years earlier. I commented that, given the quantal nature of advance, this was to be expected. Gaines proposes that in any area of human endeavour, you first enter a *Breakthrough* stage, where attempts are made to solve an unsolved problem with many failures, false starts and little progress, because there is no experience, let alone theory or even methodology. With aeroplanes, for example, all kinds of ideas were tried. Films of early attempts make entertaining viewing. With hindsight, it is too easy to ridicule the failures. Eventually someone

---

1. One topic that came up at our conference illustrates this point in a minor but important way. In trying to improve the cockpit of military aircraft engaged on low-level missions, researchers at the Applied Psychology Unit in Cambridge, England, have addressed the problems of warning pilots of urgent problems with startling them, since being startled at low altitude could prove fatal! This has led to some very interesting research on noises which are difficult to overlook, but do not startle.

succeeds; the Wright brothers managed to get a piece of equipment aloft for a while at Kittyhawk, despite knowing nothing about the theory of aerofoils, control methods, or very much else. But they had abstracted and implemented just enough to succeed.
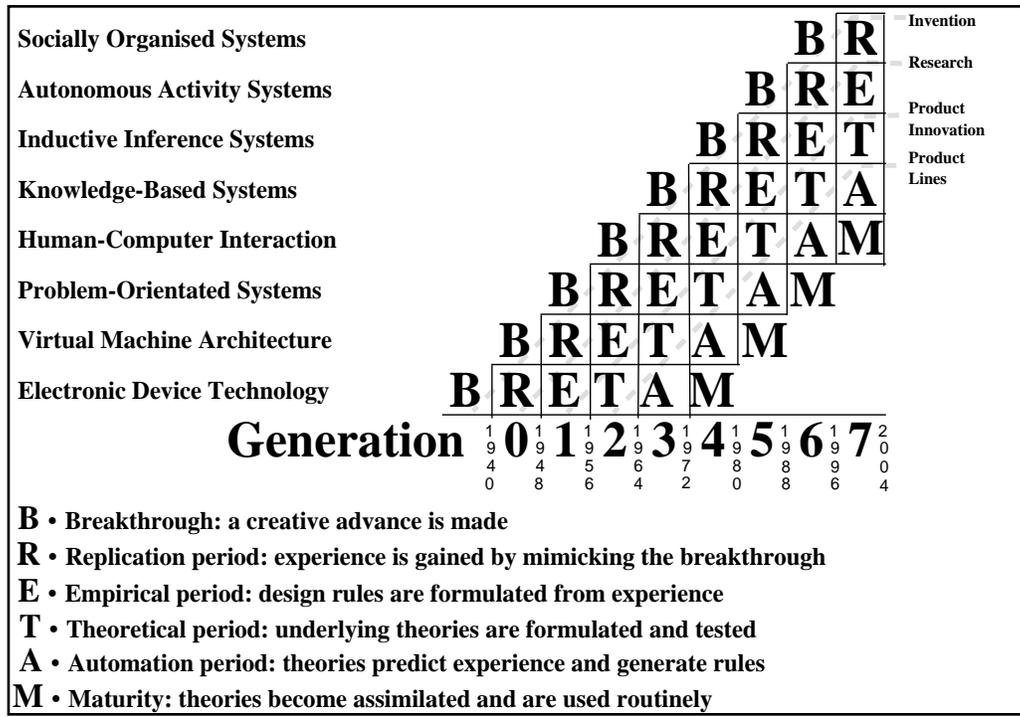
		Invention
Socially Organised Systems	**B R**	
Autonomous Activity Systems	**B R E**	Research
Inductive Inference Systems	**B R E T**	Product Innovation
Knowledge-Based Systems	**B R E T A**	Product Lines
Human-Computer Interaction	**B R E T A M**	
Problem-Orientated Systems	**B R E T A M**	
Virtual Machine Architecture	**B R E T A M**	
Electronic Device Technology	**B R E T A M**	

Generation **0** **1** **2** **3** **4** **5** **6** **7**
1940 1948 1956 1964 1972 1980 1988 1996 2004

B • **Breakthrough: a creative advance is made**
R • **Replication period: experience is gained by mimicking the breakthrough**
E • **Empirical period: design rules are formulated from experience**
T • **Theoretical period: underlying theories are formulated and tested**
A • **Automation period: theories predict experience and generate rules**
M • **Maturity: theories become assimilated and are used routinely**

Figure 1: Gaines & Shaw's model of technological/cultural progress

The importance of such an unstructured breakthrough stage as a precursor of innovation and cultural advance is built right into the credo of the *Media Lab* at MIT which replaces the "Publish or Perish" of conventional academe which has connotations of learned discourse and theory, with its own "Demo or Die" (Brand 1987, p 4)—reminiscent of the ferment of early aviation wannabes.

Once someone has demonstrated success in a new enterprise (shades of Machiavelli), people see it, admire it, and copy what was done, usually without understanding why it works. Almost certainly they inadvertently introduce variations which sometimes improve on the original and sometimes do the opposite. This is the *Replication* stage which provides experience and the beginnings of empirical rules, if not theory. With enough experience, the lessons learned provide a basis for the *Empirical* stage of new advance, where experience is categorised into experimental rules of thumb that are still fairly superficial (what works, what improves performance, what degrades performance, and so on). Practitioners have some established basis for continued success. This in turn provides the source material for abstraction by theorists and formal experimentalists—hypotheses are formed concerning the laws governing the phenomena of interest, and experiments or other tests are designed to test these causal theories. This is the *Theory* stage which leads to formal understanding, or at least excellent predictive models. Once the new phenomena are sufficiently understood, they can be designed into new

structures and artifacts with assurance of success or, in the case of information processing, they can be automated—the *Automation* stage. Finally the new technology or cultural scheme becomes thoroughly assimilated and is used almost without thinking, and has reached *Maturity*. Like a thunderstorm, maturity may contain the seeds of dissipation as well as a new thunderstorm.

Thus, more dramatic than the evolution of a particular technology like electronic devices (represented by the horizontal dimension of Figure 1) is the evolution to new technologies which appears as the vertical dimension (specific to progress in computing in this case). The experience with a new technology leads to new creative advances. The ability to store and modify programs electronically allows the basic hardware capabilities of the early computing devices to be extended. Code segments provide a variety of virtual machine possibilities, and provides the foundation of the new technology of Virtual Machine Architectures, distinct from the initial technology of electron devices. The new technology breakthroughs depicted in Figure 1 correspond to the acknowledged generations of computers, with some extrapolation to what lies in the immediate future. A given stage of technology advance takes roughly eight years. New generations have appeared at roughly eight year intervals, whilst the delay between invention and product innovation is roughly sixteen years, with a further eight years for product innovation to become established practice. We are still looking for breakthroughs in managing the new complexity of the computer human interface.

1.4 Balance and the complexity imperative

Interestingly, and for his own reasons in the context of information technology forecasting, Gaines quotes Luhman (1979) and De Bono (1979) to support the view that the fundamental motivation for all our social institutions is *complexity-reduction*, as part of the process of self-preservation and progress.

> *The world is overwhelmingly complex for every kind of real system... Its possibilities exceed those to which the system has the capacity to respond. ... Man has the capacity to comprehend the world, can see alternatives, possibilities, can realise his own ignorance, and can perceive himself as one who must make decisions. (Luhman 1979)*

> *By great good fortune, and just in time, we have to hand a device that can rescue us from the mass of complexity. That device is the computer. The computer will be to the organization revolution what steam power was to the industrial revolution. ... Of course we have to ensure that the result is more human rather than less human. Similarly we have to use the computer to reduce complexity rather than to increase complexity, by making it possible to cope with increased complexity (De Bono 1979)*

Neglecting the Old Testament, controlling, politically incorrect, sexist, saviour-myth tone of these quotations, they offer a kernel of truth and a lesson for our present purpose. I ignore the sweeping assumption underlying the whole set of

arguments that progress is both good and inevitable, and that the destiny of the world must be managed by proactive human effort.

While computers enormously extend our organising power and ability to manage complexity, they also provide a novel means of increasing the complexity with which we have to cope. We must tread a narrow, delicately chosen path that matches our ability to manage complexity with our ability to create it. It is these twin rocks, the Scylla and Charybdis of the information age, that provide the basic context for Phase 3 of CHI. The very tool that offers salvation through modelling and managing the complexity of the world, may itself provide enough overwhelming complexity to defeat its purpose. Whatever we may think, we are still in the breakthrough stage in trying to manage the huge variety of interface tools that are, and will become available.

Today's *hot* areas reflect moves to satisfy these needs. Computer-supported Co-operative Work and Groupware addresses the need to manage the interaction when more than one or two people are involved. Programmer's Workbenches and Adaptive Workspaces address the need to manage the tool-sets needed for different task contexts. Knowledge Elicitation addresses the need to manage the acquisition of knowledge to support expert system behaviour. And so on. Such tools reside at a higher level in the layers of organisation supporting interaction.

Eventually, to solve the problem, I suspect we are up against that ultimate philosophical conundrum: we must understand ourselves.

2. The challenge of managing the tools

2.1 The context

Humans have always been toolmakers. Whereas Phase 1 mainly addressed the problem of making complex physical tools more "available", more productive and less dangerous, Phase 2 was concerned to create new tools for the intellect—information processing tools such as the mouse, the graphical tools first seen in *Sketchpad* (Sutherland 1963), the object-oriented paradigm, convenient direct-manipulation editors, fish-eye views, font editors, animation systems, and speech input/output, to name only a few. Phase 2 also had to deal with tool availability, productivity and reliability problems at this cognitive level. An important part of this research was tapping into the skills and mental structures of users in order to match the interaction to the users' methods, concepts and abilities, so that some of the progress was conceptual, touching on Piaget and Bruner's work on mentalities, and the need to address appropriate "mentalities" directly. Alan Kay's 1987 video (Kay 1987) provides a penetrating insight into the ferment of these years.

We have now reached the point where there is a huge variety of tools, of conceptual bases for interaction, of modalities for action, and so on. As I have argued above, the challenge now is managing this complexity.

The complexity arises partly from the number and variety of tools themselves, partly from the contexts in which the tools are used, and partly from the need to enrich and extend the interaction. It is increasingly difficult to use the most fundamental principle of interface design—"Know the user" (Hansen 1971)by itself. Knowing the user means evaluating and using the human context and goals in relation to which a particular computer-aided task will be carried out. Until now, that principle has been a good basis for UI designers to make appropriate decisions concerning conceptual design, tool and model selection, task allocation, functionality, presentation, learnability, error management, physical and psychological compatibility, and acceptability. Most design guidelines flow from it.

In Phase 3, with so many options, so much variety, so many possibilities for metaphor and analogy, and such increasingly ambitious designs, the user needs more than the availability of a set of good tools and stock procedures geared to existing mental models. The user needs assistance in managing the interaction, in the moment to moment selection of tools, and in navigating through the particular virtual reality, in a context comprising a multiplicity of tasks, channels of communication, possible views and modes of thought, and modes of interaction.

2.2 Modes, medium (media), multichannel, multilayered protocols and related "m" word topics

This subsection is intended as preparation for some points I make below about models for communication, particularly those proposed by Rasmussen (1983) and by Thomas (1978) and Thomas & Carroll (1981)

It is surprising that at a workshop on the structure of multimodal dialogue we seemed to disagree so much about the meaning of terms like mode, medium and channel. While doubting that clear definitions are necessary, Pierre Falzon takes *mode* as something defined by a lexicon, and a syntax, having particular semantics and pragmatics, supported by a *medium*; whilst a medium is a physical capability related to sensors and effectors (of a human or machine partner). Thus, in allocating a mode for purposes of dialogue, one may choose vision as a medium and then choose natural language, or graphics, as the mode. Or, having decided to use natural language, you could use the medium of vision (with output on paper or a screen) or the medium of speech. Understanding modes, media and the like is important in the context of dialogue structure and management.

There then arises the question as to whether all modes behave in the same way with respect to their expressive power. Falzon gave saying "Good morning" or proffering an outstretched hand as being equivalent in both expressive and reactive power. In either case, the recipient of the communication is pretty well compelled to react in a particular way. Sylvia Candelaria de Ram pointed out that the handshake was far more compelling than the verbal greeting, but that many people would simply not know how to respond to an outstretched hand at all (being culturally determined). She suggested that what was mode and what was medium was not clear (at least in this case). Daniel Luzzati then pointed out that even when using vision [graphics] as the mode, natural language may well mediate

interpretation, while Mark Maybury noted that mode allocation will likely depend on the content to be conveyed, with graphics suiting some purposes and natural language others.

I feel that these kinds of arguments miss some fundamental points about the nature of mode and medium, and their relevance to the communication. If you take Rassmussen's *abstraction hierarchy* (Rassmussen 1983) in dealing with CHI, then if you choose the medium, you are working bottom up, while if you choose mode you are working top down. *Mode* is associated with *purpose* whereas *medium* is associated with *means*. In the slide Pierre Falzon used as one illustration, he showed a pie-chart and a histogram with the same information. Now the pie-chart and the histogram actually implied different purposes, while the choice of paper (the *medium*) did not imply any particular purpose, but could have allowed many. Paper was simply the means, the *physical particulars* according to Rasmussen, not specific to the *functional purpose*.

When discussing communication, people often use the term *communication channel*. Shannon used the term channel in a formal sense, to mean a physical means (a *medium*) used in a particular way, with no assignment of purpose. The notion of "particular way" is now subsumed under the term *protocol*, and multilayered protocols became a topic of some importance at the workshop in the context of some of Martin Taylor's work. If you agree with Thomas and Carroll as I do (see below), you will reject layered protocols as a basis for understanding human computer interaction except perhaps at the very mechanical level of how to structure media on the physical (syntactic) level. They prefer the Design-Interpret (DI) model of dialogue over the Encode-Decode (ED) model. In the DI the goals of the participants continually shape the form and content of the dialogue. This is really the point made by Pierre Falzon when he noted that the mother in William Edmonson's example (this volume, Chapter 18) was trying to fulfill a variety of goals (teaching, avoiding aggression, avoiding direct orders, …). Taylor and Waugh, this volume, Chapter 25, present an opposing view. Martin Taylor and I simply disagree, though it may be very instructive to see how far you can take the layered protocol stuff in the context of dialogue, turn-taking, understanding and purpose. It may be a useful way of finding out what we don't know, but is surely bottom up.

Shannon restricted his attention to the purely syntactic level of communication. Attempts to extend the ideas to higher levels (semantics, pragmatics) have not been successful. In the present context, the term channel appears to be ambiguous because the particular way a physical means is used in dialogue necessarily relates to purpose. It is may be best to talk about *media*, and *modes* in the sense I have explained, and to leave the term *channel* unused. However, if channel is used in the context of CHI or human-human communication, it is probably synonymous with *mode* and quite distinct from Shannon's usage. If I use the term *channel* it will be in exactly this *mode*-synonymous sense.

Choosing my words rather carefully, there are really four reasons for deciding to use particular modes, more than one mode, or alternative modalities:

Multiplexing: first you may have a single medium and wish to multiplex several different kinds of information onto it—different views of the same process or data-structure, for example.

Parallelism: secondly, you may wish to monitor or take control action for several different things in parallel, which may involve different media. When you use different media, whether the modes are the same or different may be a good question. For example, it is not clear whether speech counts as a medium distinct from sound, or a mode distinct from written language. I would think, contrary to some opinions that were expressed at the workshop, that speech is a mode carried by the medium of sound, and is a subclass of the mode natural language just like written language. Whether one can create an inclusive, consistent hierarchy of classes and subclasses of modes is also a good question, but I do not see any insurmountable difficulty, although memories of Foley and Wallace's difficulties with graphical input modes (1984) ought to serve as a salutary cautionary tale. It would be a worthwhile and instructive exercise.

Appropriateness: thirdly, you may want to match modes to the information conveyed (it is more effective to show a picture than to describe it). This was part of the point made by Mark Maybury, who also noted that many sources of knowledge can be brought to bear in the selection of modes, and that modes have varied characteristics. The point was reinforced by John Lee and Martin Taylor who noted that shape and texture gradient are not easily represented non-graphically, whilst abstraction may be very difficult to express graphically. In a remotely controlled bottling plant, sound may be more important than vision.

Substitution: finally you may need to substitute one medium for another. In the case of someone who is visually disabled, for example, you will need to substitute for the visual medium, and very likely you will wish to substitute *several different* modalities. This is because the visual medium is able to carry so much information that you have to split it once you start looking at touch, sound and proprioception, in order to make it manageable, relevant, and appropriate. This is one of the things we have done with the *TouchNTalk* workstation for the visually disabled, but there may be reasons other than abled-ness for substituting one set of modes by another.

Of course, these reasons are not mutually exclusive. More than one may be in operation at the same time (for a particular choice) and there may be trade-offs—not only because of conflicts arising from constraints related to the reasons above, but also because of the need to make the whole interaction manageable for the user. This point was underlined by Pierre Falzon. It is also true that particular subclasses of mode relate to particular media (speech to sound and text to vision, for example), but not all modes can be effectively carried by all media. There will be conditions under which the designer is forced to choose a particular medium or particular media, and conditions under which the designer will be forced into multimodality,

and these are not necessarily the same thing, nor are they necessarily mutually compatible.

There is also the problem, raised by David Sadek, that some kind of economic metric may be necessary. There are costs involved in choices. While we may hope that ever progressing technology will eventually remove cost limitations arising from the scale and complexity of hardware, there are other kinds of costs bound up with things like information load on the user, familiarity, clutter on the media, and response time constraints, that may influence choices. Some of these are things Mark Maybury alluded to.

Another point of importance is that modes may be mutually supportive. One mode may not be able to convey information by itself. If speech and gesture are in use, a verbal command will likely only make sense if it can be related to a relevant gesture. If speech is in use, perhaps its has submodes. Tom Wachtel argued that some modes are more multi than others. Natural language by graphics—strings of characters—is a very underprivileged mode of natural language compared to speech, where intonation and rhythm, amongst other things adds considerably to the meaning. The way you ask a question can greatly affect the answer you receive, but this involves pragmatics as well as semantics. The difficulties of managing such aspects are non-trivial, and require the machine system to use fairly sophisticated models of the task, the language and the user. Computer speech output may make little sense if the computer is unable to understand what it is saying, and use this understanding to apply an appropriate rhythm and intonation models correctly. Whether more than one mode is involved in this sort of coordination is another good question, but intonation, by itself, can convey significant information. And then there's lip reading. Is visual speech a separate modality? Adding a face to computer speech output certainly should increase the intelligibility, if done correctly. These are other lines I have started researching. Answers to such questions are necessary for progress in multimodal CHI, but are clearly basic to AI as well as other disciplines. In human-human communication, a great deal is communicated without easy conscious control, and different things may be communicated simultaneously (for example, the message conveyed by the sounds of speech may be at variance with that conveyed by rhythm and intonation, or body language). Whether it is appropriate to try and capture such emotional/attitudinal aspects of communication in CHI at this stage is another good question, but clearly an automotive computer might find some of the information useful for thwarting a drunk driver, especially if coupled with a sense of smell!

One final point should be raised, that also got mention at the conference. When humans communicate, they do so according to their cultural background, and this will, amongst other things, include a range from high context to low context. E.T. Hall (1981) discusses the difference between high context communication, in which much is taken for granted, and low context communication where everything has to be spelt out. The Japanese swing wildly between extremes. In high context mode you are expected to know that it is an honour to be moved from room to room without notice or permission, when staying at a hotel. It is part of the culture that

you are being treated like a family member, you *belong*. At the other extreme, you may have to state explicitly that your brown shoes are to be cleaned with brown shoe polish (rather than some other colour) (Hall, 1981, p.66). People tend to use a higher context style of communication with each other than they would with a machine, and a theatre ticket salesperson, for example, must not respond literally to the question "When is the next performance?" if it is already sold out. It will be profitable to move machine dialogue towards a higher level of context, which is the issue involved in Tom Stewart's "set-breaking" example (Section 3.3.2). Equally, there are acceptable ways and unacceptable ways of communicating. It may ultimately be counterproductive to tell even your own child that it is naughty to pull the dog's tail, although William Edmonson's suggestion "Johnny, I am sure that Fido likes having his tail pulled" goes many layers beyond the more reasonable "If you had a tail, would you like someone to pull it?"

Understanding ourselves presents formidable problems, but progress in CHI demands that we travel that path, the key to which is partly buried in certain kinds of AI research involving expertise and understanding.

3. The formation and use of models: expertise at the interface

3.1 Some more "m" words: models & mentalities in relation to knowledge.

Model-based activity is central to *human* perceptual, communicative and problem-solving processes (Gentner & Stevens 1983). Rissland (1984) raises essentially the same point when she talks about the importance of *sources of knowledge*. Knowledge and models can exist at various levels of abstraction, and may be built up in a hierarchy. Models represent sources of knowledge for planning, testing and evaluating in knowledge-based performance. Models can represent the generative core knowledge that allows a user to avoid rote memorisation of procedures in dealing with complex systems, reconstructing them instead and even generating new ones (Halasz & Moran 1983). Other kinds of models represent the user's understanding of the system, the task, and methods of interaction.

In an expert system, expertise (knowledge) is the prime commodity being accessed. An expert system *models* part of the real world in a form that allows the knowledge to be used for prediction, problem solving, and/or data access.   The model encapsulates knowledge in a usable form. This kind of modelling is one of the AI-hard problems. A true expert system must be able to explain itself to its user. This requires the creation and use of additional models (of the task, of the system behaviour, of the user, and so on) to drive the interaction and to allow the necessary explanations to be provided. These models, and the routines that use them, are no different in kind from the task-determined expertise, but they represent knowledge the system has about itself, or the user, or the context, or the available

communication channels—beyond task-determined expertise. Together, they may be thought of as a limited form of self-consciousness that is essential for excellent interaction. Models (knowledge sources) are needed to support the task and the models (knowledge sources) are needed to support the interaction.

It is important to remember that both participants in a dialogue use models. It is not enough that the machine be able to model the real world, including the user. The machine is an important part of the user's real world, and must itself be capable of being modelled by the user. This is why user interface guidelines require uniformity and consistency in the interface, but that is only a passive aid to helping the user in this part of his task. The machine should actively reveal itself to the user. Some of the models relevant to computer-human interaction must exist within the machine, and some within the user, but, in a very real sense, all these models must be shared. Even inaccessible models must be shared in some form.

Thomas Sebeok, the well known linguist, conjectured that the ability to model the real world was developed by humans prior to the use of language, and provided the basis for the development of language, which came two million years later (Sebeok 1986). Language is a way of sharing (and updating) models. Whether you accept Sebeok's conjecture or not, it is clear that modelling is an important part of intelligent behaviour, whichever side of a dialogue or interface is under discussion.

Models, as I am using the term, are at least partly procedural in the sense that they have a degree of autonomy and may be used for the prediction of consequences or the generation of other sorts of information by emulating (more or less accurately) the behaviour of some part of the real world. They are fundamental to knowledge representation, problem solving and language. Only by mastering the creation and use of models in our machines, and by developing machines that help us create and manage richer, more accurate models in ourselves, shall we solve some of the hard problems that now face us. It is in these terms, for example, that I would frame the problem of Computer Assisted Learning (CAL)—still largely unsolved, and a hot target for the "multimedia crowd" as I write.

Interestingly, this line of reasoning brings us to the topic of Piaget's *mentalities*, raised in Alan Kay's video presentation (Kay 1987). There is a great tendency to assume that all reasoning is abstract/symbolic, and that beyond the age of fifteen, all thinking humans learn best symbolically. Kay demolishes this idea, and shows that different mentalities require different approaches to instruction (model building), and more than one mentality may be in operation at the same time. It would be as well to bear this in mind as we remarry AI and CHI, for AI is already deeply wedded to the use of abstract reasoning techniques, yet we know that people like Einstein used other mentalities for their creativity, and only resorted to a symbolic mentality for communication.

3.2 Managing complexity using knowledge: models in action—early examples

A metaphor is a model of some aspect of reality that is useful in

understanding some other aspect. The Macintosh uses the desktop metaphor as a way of understanding how to access and use procedures and data represented as icons on its CRT display. In this way, complexity is controlled by being rendered familiar. Carroll & Thomas (1982) and Thomas & Carroll (1981) provide excellent coverage of the importance and use of metaphors. Interaction is more effective when the models and metaphors are close to the reality they mirror, unless the task is strictly routine (i.e. rule-based—see the next section) (Halasz & Moran 1983).

A system can actively help the user to form correct models of itself or suggest appropriate metaphors. The user can contribute to the system's models of the user and the current task. In RABBIT, a database retrieval interface (Williams 1984), this is the role of "reformulation", a process of interactively refining the original query to match the target(s) sought. An important function in computer-human interaction is to update and correct the relevant models. The process involves both tutoring and knowledge elicitation. Rich (1983) provides an excellent study of the system modelling the user.

The success and power of SOPHIE (Brown 1975) as an instructional system for electronics depended in part on the excellence of its modelling and reasoning, and in part on its ability to communicate using language in a natural and robust manner, both depending on a great deal of built in knowledge (circuit simulation, natural language parsing, spelling correction, …). The modification of the circuit model to represent faults, and the reconciliation of the fault model with observation, gives an early example of the kind of knowledge-based interaction that will come to dominate future CHI.

3.3 Managing complexity by understanding

3.3.1 RASMUSSEN'S ABSTRACTION HIERARCHY AND MODEL VALIDATION

Rasmussen (1983) takes an in-depth look at models of human performance that can be used in the design and evaluation of new human-computer interfaces. He identifies three categories of human performance: skill based; rule based; and knowledge based. As illustration, with reference to machine performance, predictive controllers for guns operate at the skill based level, and expert systems operate at the rule based level; knowledge based behaviour lies in the domain of unsolved AI problems. It can take indefinite amounts of time to solve problems that need knowledge based performance. This was the basis for the US Navy simply dumping electronic equipment overboard if it took more than two hours to fix. After two hours, you'd exhausted rule based fixes and after that it might take forever. Knowledge-based performance requires accurate models of extensive parts of the real world to allow mental experiments to be run to predict outcomes under different assumptions and conditions. This activity may be supplemented by actual experiments on the real world to supplement internal models or create new knowledge that may be relevant. These experiments require insight and design. All this is to test theories about cause and effect and possible relationships in a problem-solving situation. In time, knowledge-based results may become abstracted to rule-based form (if this situation, take this action), and ultimately routinised to skills.

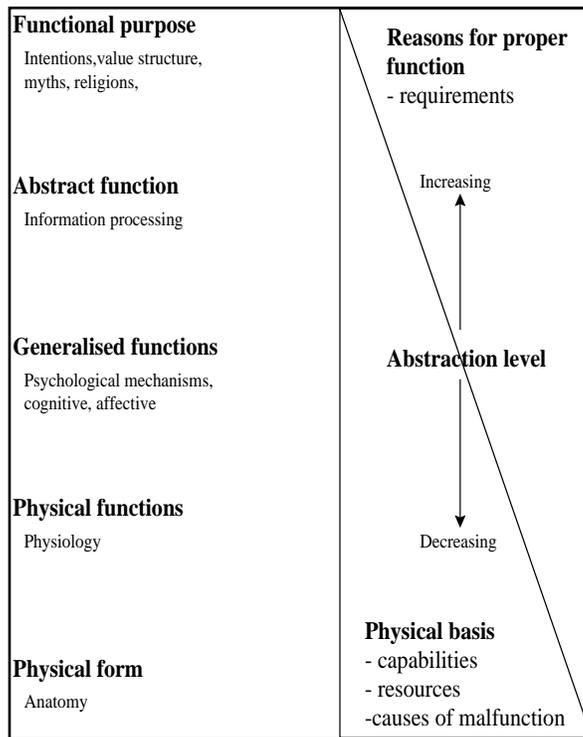Both these steps speed up performance and reduce flexibility. Such acquisition and

Functional purpose Intentions,value structure, myths, religions,	Reasons for proper function - requirements
Abstract function Information processing	Increasing
Generalised functions Psychological mechanisms, cognitive, affective	Abstraction level
Physical functions Physiology	Decreasing
Physical form Anatomy	Physical basis - capabilities - resources -causes of malfunction

Figure 2: Rasmussen's Abstraction hierarchy applied to the human

assimilation of new knowledge takes both time and considerable resources of experience (stored knowledge—i.e. other models!).

Knowledge is modelled at different levels depending on the associated behaviour (performance), and the same physical indication may be interpreted differently if the category of associated behaviour is different.

One aspect of models considered, in the context of human performance, is that the power of human cognition depends on the richness of the human's repertoire of model transformations, and the human's ability to consider abstractions of situations *at any level* as suited to the task. Changing the level of abstraction not only involves a

hierarchy of physical forms, but a shift in conception and structure as well, moving from questions of What to questions of Why. At lower levels, a given physical system can have more than one possible purpose. At higher levels, one purpose is potentially achievable using more than one physical arrangement. Model transformations at the extremes are mutually orthogonal (they don't interact) because the lowest level transformations follow material arrangements and properties, whilst the highest level follow purpose. Rasmussen's Abstraction Hierarchy, reproduced as Figure 2, diagrams these relationships.

Causes of improper function must be determined bottom up. Causes of proper function, on the other hand, are necessarily top down. Fixing a bug in a system involves moving up and down the hierarchy: top down to determine proper function (information flows and proper states); and bottom up to analyse and explain the actual state. Design also iterates amongst the levels, with the potential for many-to-one mappings either way. Design is not the orderly process sometimes claimed. Rasmussen says the *physical* ("what it is") is sometimes called a *formal* description, whilst the *purposive* ("what it does") is called a *functional* description, quoting Alexander (1964). These seem closely related to *form* versus *content* to use Thomas and Carroll's terms, or *medium* versus *mode* to link it to what I said earlier. Rasmussen makes it clear that his abstraction hierarchy, running from physical

function, through generalised and abstract function, to functional purpose, involves intention (*"why* it does what it does"*) at the highest level and this is the only level that can truly be called purposive.

Errors are only meaningful in relation to purpose (intention), and errors are important in CHI. Error management is one of the most important goals of a well-designed interface. Rasmussen notes that successful performance does not validate a model. Only testing limits and error properties does that. In this he echoes the view of scientific methodology put forward by Popper (1963). I well remember my last landing before going solo when I was learning to fly as an RAF pilot. My landings had all been well executed in prior flights and this one was going to be no exception. At the critical moment, as I rounded out, my instructor rammed the control column forward, causing the aircraft to strike the ground hard, and bounce ten or fifteen feet into the air. As I cussed and juggled the throttle and stick to regain control and land safely, I wondered what on earth could have happened. When I found out, I complained. "Well," said my instructor, "I knew you were pretty hot on landings, but I wanted to know if you could safely recover from a bad landing." This is the kind of validation that Rasmussen refers to. The model must include a complete operational envelope, and the fact it works under good conditions does not validate it. If the model fails the validation tests, his methodology provides a basis for debugging and correcting it, which is an important part of managing complexity.

3.3.2 RAISING THE ABSTRACTION LEVEL AND "SET-BREAKING" (AN EXAMPLE OF "UNDERSTANDING")

Norman (1984) is also concerned that future interfaces should move away from the level of details (physical models) towards the intentional global levels. He relates the story of a man going to open his car door:

> *X leaves work and goes to his car in the parking lot. X inserts his key in the door, but the door will not open. X tries the key a second time: it still doesn't work. Puzzled, X reverses the key, then examines all the keys on the key ring to see if the correct key is being used. X then tries once more, walks around to the other door of the car to try yet again. In walking around, X notes that this is the incorrect car. X then goes to his own car and unlocks the door without difficulty.*

He reports having a collection of stories similar to this, revealing that even though people know their own intentions, they seem to work bottom up, tackling the problem at the lowest level, and only reluctantly and slowly moving to the higher levels of action and intention [quite the reverse of dealing with people]. There is a role here for an interactive system to prod users out of inappropriate levels, and away from incorrect hypotheses. Norman asks: "What if the door could have said 'This key is for a different car'"?

It has been established that one common failure mode in human problem solving is failure to abandon an initial hypothesis. In one study (Wason 1971), students were asked to determine the rule underlying the generation of a number sequence, given the beginning of the sequence. If the rule was incorrect, further

numbers in the sequence were given, refuting the initial hypothesis, and providing more data. Many of the students simply reformulated the original, incorrect hypothesis, perpetuating their inability to solve the problem. This behaviour is seen in Norman's example. T.F.M. Stewart has called this the "set-breaking" problem.

Clearly, in directing users away from incorrect hypotheses, there is a role for machine understanding and modelling intentions—good Phase 3 stuff.

3.4 Managing complexity by dialogue design: Design-Interpret versus Encode-Decode—appropriate models for design

It seems to be generally accepted as a design principle that computer-human dialogue is closely related to human-human dialogue although there are those who feel uncomfortable with the idea of trying to mimic human-human dialogue as a basis for CHI. Even when CHI dialogue is totally rudimentary, people tend to anthropomorphise the machine (note how some people treat their automobiles!). One problem with sophisticated dialogues is that people may be misled as to the abilities of the machine, so that failure to respond in a human way in certain situations can be very disruptive. Taylor (1988) has described a layered protocol model as a basis for understanding and designing dialogues in the context of CHI. The model seems strongly Encode-Decode (ED) oriented, to use Thomas' formulation (Thomas 1978; Thomas & Carroll 1981). Thomas and Carroll wax eloquent about an alternate model for dialogue which they term Design-Interpret (DI) in which messages are designed to have a particular effect, given the context which includes the responses and goals (or purpose) of the participants. In Rasmussen's terms, the big difference is that the ED model emphasises the physical and ignores the purpose whilst the DI model concentrates on exactly the functional purpose. In communication theory terms, the ED model seems to ignore the uncertainties concerning the receiver's state, and the need for feedback in dialogue. The ED model assumes that a message can be complete in itself, and any failure is due to a fault in the receiver, rather than a consequence of the circumstances or context of communication, such as conflicting goals, differing experience (model sets) of the participants, differing task requirements, and so on. The DI model is adaptive, and takes such factors into account automatically because it is driven by function or purpose and directs the allocation of dialogue resources to fulfill it.

3.5 Selecting modes, sharing goals

In formulating the structure of multi-modal dialogue, Rasmussen's abstraction hierarchy and the DI model of communication are powerful tools for coming to grips with the requirements and designing solutions at all levels. One important basic idea is that, in CHI, the computer should massage any data that are presented into a form that matches the category of behaviour used for the task. This is one basis for selecting *mode*—a topic that was well aired at the workshop.

Rasmussen states that people use almost exclusively top down models of other people in dealing with them [rather than the bottom up approach they seem to

adopt with machines, according to Norman [2]]. People base their behaviour and responses on perceived intentions, motivations and capabilities. Causal (bottom up) explanations play little part.

> *… the most important information to use for planning human interactions for unfamiliar occasions is therefore the value structures and myths of the work environment. The obvious reason for this is the complexity and flexibility of the human organism. However, it should be emphasized that due to the growing complexity of information and control systems, the role of such intentional models (Dennett 1971) is rapidly increasing, and for interaction with such technical systems as well. (Rasmussen 1983, p 263)*

There is a need for research into this area in terms of CHI interactions, for it is not clear that people ascribe intentions to mechanisms (but, again, consider how people react to their automobiles). It is obvious that a dialogue is likely to be more productive if both partners understand the purpose of the discourse or common task (witness Norman's car key problem). It may be helpful to consider and catalogue the system goals, when designing interaction, and be sure that these goals (the machine's *purpose*) are clearly revealed to the user. After all, it is considered important to state objectives clearly when lecturing, which is simply making the purpose of the lecture explicit. Equally, the machine needs to be tuned to the user's goals, and perhaps needs to maintain the user's focus on these goals as well as support them.

4. Redundant, multimodal communication; pointing, looking and situational cues

As noted in the introduction and at the workshop, interaction modes may be mutually supporting. The issue is raised explicitly in a chapter entitled "Future Interfaces" in Bolt's book about the work of the *Architecture Machine Group* (Bolt 1984) [3]. Noting that related information in a communication channel may be redundant or supplementary in character, and that this form of communication was invented by nature, he points out the advantages of being able to speak, point and look, all at the same time (the *AMG's Dataland* allows all these modes to be sensed [Bolt 1980]). Supplementary information supports such possibilities as the resolution of pronouns by pointing (making speech more economical and natural).

---

2. In terms of Rasmussen's abstraction hierarchy, one might explain the difference in the way that people react to machines as opposed to other people by suggesting that they are very often effectively dealing with malfunctions when dealing with machines, so that a bottom up approach is appropriate. When dealing with other people, the focus is usually on communication, when purpose and intention become dominant, so that the top down approach is then preferred.
3. The Architecture Machine Group has now evolved to become, not surprisingly, the Media Lab.

Redundant information can ensure correct identification of intent from information that, taken piecemeal, is ambiguous due to imperfections of various kinds. Bolt also emphasises the importance of integration of sources of information. The usefulness of the whole is greater than the sum of its parts.

It is important to realise that this kind of integration can only be based on relevant expertise that "knows about", and can manage related information entering via different channels. This is another area needing considerable research on both theory and implementation.

The whole concept of the *AMG's Dataland* is futuristic. However, in summarising the "Future Interfaces" chapter, Bolt picks out two other main points as especially relevant. One is the use of circumstantial cues, particularly in retrieval, and the other is the use of eye tracking to help in modelling the user's interests and intentions—more good Phase 3 stuff.

We tend to remember circumstances, even when contents are forgotten, and even though the circumstances may have little formal connection with the desired fact or action. Thus we remember information from books, reports and newspapers partly on the basis of where and when we obtained it, as well as whereabouts within the source the information was encountered. Such information provides hooks to access our associative memory, and explains why examinations seem easier if they are taken in the lecture room, rather than some new place. Far from preserving this kind of information, current computer systems even suppress what little may exist. Thus text, presented on screens, is likely to change in format, depending on the terminal, or the trivial modifications made since a previous session, whilst failing to give any one document distinct visual character. A rich CHI system can keep the circumstantial record of a user's activities and preserve the distinct idiosyncratic form of documents even in electronic form, using such cues to assist in future interactions. It is possible and probably reasonable to envisage a system in which "printed" material could only be viewed in bit-mapped run-off form, which could also help in copyright protection (Benest and Jones 1982) [4]. In passing, it is worth noting that one of the major complaints that users make about the Apple *Newton*—a Personal Digital Assistant (PDA, or pocket computer-*cum*-memo-book), with a handprinted character recognition input), is that the original form of the input is lost when the *ink* laid by the user is translated to text characters. Not only does this destroy the visual aspects of any notes made but, with far from perfect recognition, the resulting text may be unrecognisable to the person who wrote it. This limitation is one of memory capacity and not all PDAs suffer from it, though there are (of course) trade-offs.

Cheaper approaches to structured document viewing that tie in with document preparation are also possible (Witten & Bramwell 1985). If we think of the different cues involved (chapters, sections, tables, fonts, bookmarks, position in the document or on the page, tea-stains, marginal notes, and so on) as communication modes, we begin to appreciate the real essence of multimodal computer-human interaction.

Cues derived from eyes, Bolt notes, are especially revealing (Bolt 1982; 1984). They

---

4. Of course, pirates could easily run an optical character recognition algorithm on such material, but it would make casual pilfering a little harder, and ensure that no pirate could claim innocence when caught.

form a highly mobile pointer, revealing interest and focus of attention, as a supplement to dialogue. A great deal can be communicated by a changing point of regard, both intentionally and unintentionally. A child learns the names of things by hearing the names and noticing what the namer is looking at. Bolt distinguishes three kinds of looking: *spontaneous*; *task-relevant*; and *changing orientation of thought*. In addition, there are pupil size effects that relate to degree of interest as well as stage of task completion. Thus, although there is clearly a need for more research, it is in principle possible to determine what a user wishes to know about, how interested the user is, and how the user's mental tasks are progressing—especially when information from the eyes is coupled with other cues like voice and gesture. This, and other multi-modal input can be used to form and update appropriate models related to the overall management of the interface.

5 Managing the interface: control, specification and prototyping

The control and management of human-computer interaction in Phase 3 CHI systems will depend on the success of research now in progress. An early review of progress, which is still valuable and up-to-date, appears in Pfaff (1985). Attention is focussed on: the functional divisions within the overall User Interface Management System (UIMS); on the location of control (should it reside in the application, or in the UIMS); and on the nature of the method(s) used for formal specification of the dialogues that are the object of the UIMS.

The UIMS, which mediates between a user and an application, is intended to provide a framework for the construction and run-time management of user interfaces, cutting out repeated hand coding of common parts. The UIMS also allows complexity management; and provides uniformity, consistency and other desirable properties in the resulting interface as a result of the constraints and facilities it embodies. Given a means of formal specification, it allows certain kinds of error and interaction performance to be verified. An excellent early start on this line of research was made by Kieras and Polson (1984). By representing the task structure and (using techniques from cognitive psychology) the complexity of knowledge needed by the user to accomplish the task, the relationship can be clarified, problems identified, and the task structure modified to improve the interaction.

Finally, a properly constructed UIMS allows interactive systems to be *prototyped* very rapidly, with user involvement and feedback. This is so important in practical applications that one should really talk about User Interface and Prototyping Management Systems (UIPMS's).

Significantly, any UIPMS *will require a consistent interface of its own*, which would provide a uniform basis for interaction for all designers. Like software for other computer methods, the UIPMS itself would be much easier to design and implement if it were already available to assist in the task, but a bootstrapping

approach will have to suffice, given the framework. The ultimate development of the idea would conceivably eradicate the distinction between programmer and non-programmer by making problem solving (CHI) and/or the definition of problem solving methods (designing CHI for some task) effective, productive and fun for anyone with a problem to solve and access to a computer. That was certainly Smith's ideal (Smith 1975). But then, that is what the inventors of FORTRAN hoped.

6. TouchNTalk: an example of multi-modal CHI

6.1 Introduction

The *TouchNTalk* workstation for visually disabled users is based on the original NeXT workstations and cubes. The current β-prototype: is based on work carried out in the author's lab at the University of Calgary; is now being developed commercially by the author's technology transfer spin-off company, *Trillium Sound Research Inc.* and will eventually be ported to other *NEXTSTEP* platforms

.

Function selection
Paragraph holophrast
Sentence holophrast
Phrase holophrast
Line/column holophrast

Horizontal scrolling
Chapter/bookmark holophrast
Find/change position in volume

Page back

Page forward

System Info Line   Main Window   Vertical scrolling

Find cursor
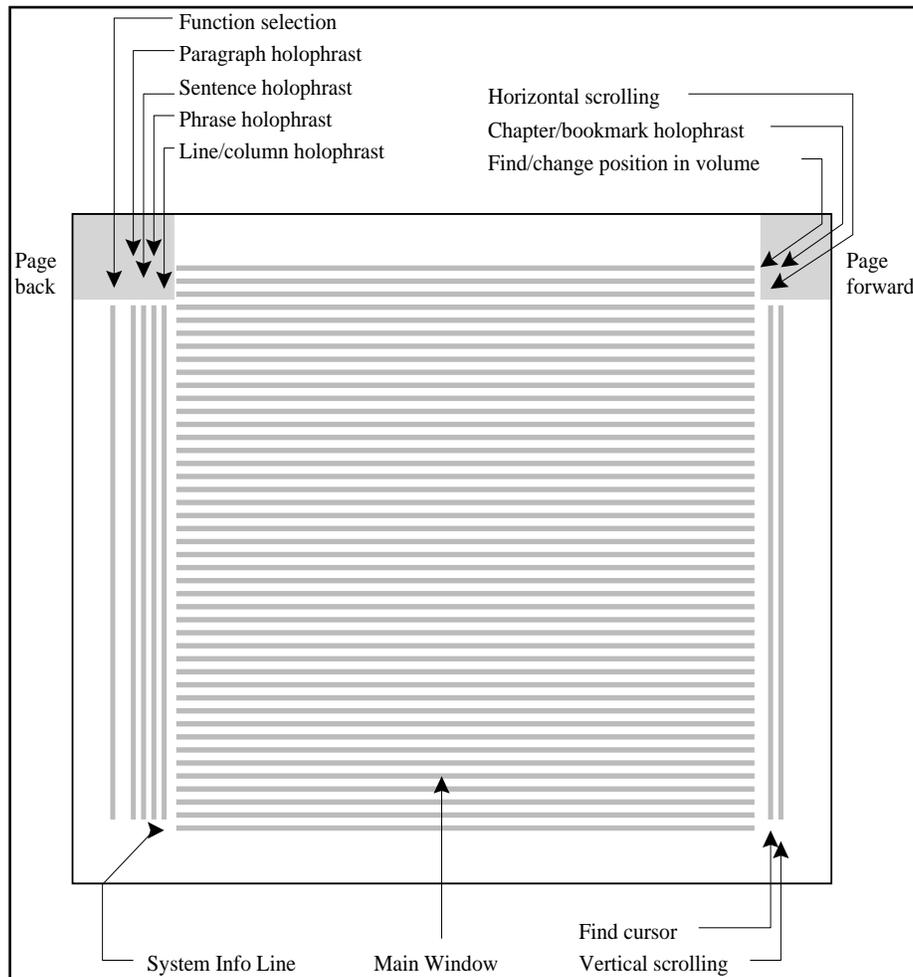
Figure 3: Diagram of the TouchNTalk textured surface

To provide the various forms of communication needed to carry out its functions, a speech server and a textured digitising pad are incorporated into the system. Other accessories such as a modem, Braille printer or Versabrailler could be added. Figure 3 shows a diagram of the textured pad currently in use. The pad, which may vary in design, is central to the system, providing a tactile/ proprioceptive model of the screen that, coupled with text-to-speech translation, becomes a pseudo-display. The concepts, methods, and tools were tested using an earlier prototype system (Hill & Grieb 1988).

6.2 Speech output

Speech output is produced by a Text-to-Speech object that exploits the built in facilities of the *NeXT*—particularly the 56001 Digital Signal Processing chip (or DSP), CD quality sound output, and interface building software. A real-time formant synthesiser runs on the DSP chip and is driven by parameters synthesised by rule from discrete phonetic text. The phonetic text, in turn, is derived from ordinary text using full dictionary lookup backed by a parser to handle special items such as abbreviations, dates, fractions and decimals, together with letter-to-sound rules for words that are not in the main dictionary. This approach is possible now because of the dramatic decline in price of main memory chips and large, fast discs. The approach has obvious advantages over morph decomposition and pure letter-to-sound rules, including the fact that it makes stress and parts-of-speech information directly available. Supplementary dictionaries for user and applications specific items are provided in addition to the main dictionary, and interactive tools are included to maintain these. Models of English rhythm and intonation are incorporated, but are presently driven only by punctuation. Variations in rhythm, intonation and speaker characteristics are possible (Hill, Schock & Manzara 1992; Manzara & Hill 1992; Taube-Schock 1993).

6.3 The pseudo-display

The current textured pad is 12 inches square and has a central portion about 11.5 inches by 9 inches laid out with 44 horizontal grooves (rows) in a pattern analogous to the lines of text or Braille in a book. These are flanked by vertical grooves (columns)—5 on the left and 2 on the right. No specific fixed characters are represented in the grooves, or anywhere else on the pad. However, the embossed rows and columns, and other surface features, may be felt to help in keeping track of position and movement when using the digitising stylus on the pad. The pad provides a physical reference for what may be thought of as a virtual screen on which text and function selections may be displayed and/or activated. The pad, backed by synthesised speech, thus forms a pseudo-display.

In our original testing, it turned out that, contrary to expectation, a stylus was preferred to the finger for indicating location because it was more precise, and the finger could still pick up tactile cues while using the stylus.

The real screen (CRT display) is also made to show what is placed on the virtual screen, partly to help in software development for the system by sighted

programmers, and partly to allow sighted colleagues to collaborate with visually impaired users.

6.4 Organisation, cursors, and controls

The particular material referenced at any given time is determined by the computer and user working together. The computer is able to detect gestures (rates and kinds of movement) as well as position, so that a variety of flexible, convenient forms of control and pseudo-display are possible.

Of the 44 rows on the tablet, 41 represent a "window" that reveals a portion of a file or document. Row 44 is the System Information Line (SIL). It retains a record of the last system activity and can be read just like any other line. Rows 4 through 43 provide the window. The characters that are revealed in the window are placed in simple one-to-one correspondence with distinct physical co-ordinates on the textured pad. Moving the stylus along the horizontal grooves causes the system to speak the corresponding text. The first 3 rows provide other special facilities as described in Section 6.7.

At any given time, a particular character within the material being dealt with is remembered by the computer as its working location. A special flag, called the *system cursor*, marks this position so that the user can also find it.

The user's working location, corresponding to the point of regard or eye fixation point for normally sighted users, is marked by a dynamic cursor, the *user cursor*, so that the system can keep track of it. The system and user cursors may or may not coincide, depending on what operations are taking place. It is always possible to move the system cursor to the user cursor, or to find the system cursor. In addition, a secondary system cursor called the *mark*, can be set to keep a memory of another location.

In addition to setting up a main working area, provision is made for natural convenient access, function selection, and control. Two columns run from the top edge of the pad to the bottom, on the right-hand side. One of these allows the system cursor to be found or manipulated quickly, by nulling an audio tone as the location of the cursor is approached (see Section 6.6).

A double tap at the location of the system cursor in the window will exchange the mark with the system cursor. The swapping process may be repeated indefinitely, but provides a means of locating the mark, if this should prove necessary. With the user cursor at the system cursor, the user simply swaps the system cursor with the mark and then finds the system cursor again.

The second column at the right is used for fine vertical positioning of the window within a body of text. Stroking up or down in the column and then tapping once (to signal readiness) moves the window by an amount physically equivalent to the finger movement. If paged material is being dealt with an audible beep sounds when the top or bottom of the window reaches the top or bottom of the page, as appropriate, and the window stops. If a further stroking movement is made in the same direction, movement continues again, and a new beep signifies if and when

the middle of the window reaches the top or bottom of the page. At this point, the window cannot be moved any further using the vertical positioning mechanism. This is useful in maintaining consistency in layout whilst still allowing material crossing page boundaries to be dealt with conveniently. To move to a previous or subsequent page, the page turning mechanism must be used. If the page is turned, the window will then be positioned with its top edge coincident with the top edge of the new page. For unpaged material the right-most column simply allows continuous scrolling. Speech feedback is also provided to allow the user to monitor the effects of any actions taken.

A column on the left of the pad allows the user to access functions such as the speech mode in use (spoken versus spelled words), or opening and closing files.

6.5 Basic text access

Finger (stylus) position determines what is spoken, and its rate of movement determines the rate of speech. Above a certain rate it is assumed that the reader is skimming, and only those important words that can fit into the time available are spoken. If the user moves very slowly, it is assumed that a spelled version of the current word is needed, and the system proceeds in spelled speech mode. A function selection allows spell mode to be locked. If the user stops, the speech stops. If the user then taps at the same location, the most recent (current) word is repeated as it was last spoken (spelled or normal). If the user taps twice in quick succession (a double tap), then the system cursor is made to coincide with the current user cursor and its old position is remembered by placing the mark there. If the user moves slowly back along the reverse direction of reading, the words are repeated in reverse order (presumably some intended point has been passed). But if the user moves back quickly, it is assumed that this is a move to gesture, and no further speech or action is generated until the user cursor reaches an active function, an active holophrast node (see Section 6.7), or resumes normal movement within the text area. A similar distinction is made for movements in other directions.

6.6 Gestures

Some gestures have already been mentioned as part of basic text access (skimming, spelling, reversing and moving to). Additionally, page turning is associated with the action of stroking the top left corner of the pad to page back, or the top right corner to page forwards. This requires deliberate action, and avoids accidental page changes, acting as a gestural analog of page turning.

The results of accessing a column vary depending on just how the column was entered. To find the system cursor, as noted, a tone is nulled by moving the stylus in one of the right-hand columns. At this point, the row containing the cursor has been located. Moving into that row again elicits the tone and nulling it identifies the exact character position of the cursor. The system knows that the user is not trying to read text by the way in which the row was entered.

Other functions also depend on the recognition of such gestures, instead of

using simple buttons for explicit non-redundant function selection, partly to make selection robust, partly to avoid unnecessary "modes" and actions, and partly to tap into the normal habits of thought possessed by users.

In order to distinguish the various gestures used to access and control the system, the software is organised as a collection of autonomous "experts" each capable of recognising a particular gesture. At a higher level in the hierarchy, a series of gestures may be recognised, as in moving in the system cursor locator column and then into the text field to find the exact position. The system "knows" that the user is looking for the system cursor rather than reading the line backwards, because of the sequence of gestures. The collection of experts modularises the expertise and makes it easy to add new experts, or update old ones. An expert models expected behaviour.

6.7 Holophrasts, more special controls, and the volume analogue

A holophrastic display presents material in condensed form, as nodes, and provides a facility for expanding the nodes. Such a structure may be hierarchical. The holophrast columns (and one row) of the *TouchNTalk* operate on single level nodes and allow: chapters; bookmarks; headings/paragraphs; sentences; phrases; lines or column entries; and the like to be detected and also spoken, if desired, without having to touch the main text area (working area). The function selection allows the holophrast columns to be changed from the primary set to a secondary set to expand the range of structures that are accessible.

Four of the columns on the left-hand side of the pad provide a holophrastic view of whatever text material is being examined or edited and are called *holophrast columns*. The three top rows are also special. Row 2 provides a holophrast of chapters and bookmarks in the document, while row 1 images the size of the document/book, and where it is open, while row 3 allows horizontal scrolling.

The top row uses a tone that varies in frequency to indicate the relative size of the document. A short document will generate a tone over a short length of the row, and long one over most of it. The frequency of the tone varies according to how near the stylus is to the relative location representing the current page—the page where the document is opened. Double tapping at another location will open the document at an approximately determined new page. It may be necessary to turn a few pages to reach the exact one intended, and this is facilitated by having the page numbers echoed as they are turned.

Single tapping anywhere in the tone region of the top row will place a bookmark (which may be named manually or automatically), and will also enter the bookmark into the chapter/bookmark holophrast which is maintained in the second row.

The third row is a horizontal scroller for wider than usual texts, and uses beeps and speech to keep the user informed about the results of using it.

Moving up or down a holophrast column elicits beeps each time a relevant structure element is encountered. Different beeps mark the ends of structures, which

becomes important when dealing with (say) paragraphs that may be separated by white space, or be broken by a page boundary. When a beep is encountered, tapping will cause the unit to be read. Lifting the stylus will stop reading. When a unit ends, a beep sounds if there is another unit of the same type on the same row of the working area.

In the chapter/bookmark holophrast row, speech supplies the name of each node encountered as a supplement to the beep. Tapping at that point will open the document to the appropriate chapter or bookmark.

6.8 Host access

The workstation is intended be used as a standalone device, providing many of the facilities expected of a personal computer, with the notable exception of graphical modes. The current design allows for standalone editing but only document reading is fully implemented. Interfaces to other local software, such as spreadsheets or the "desktop", have yet to be developed. The system can be autolaunched at login, and a command line interface can be handled by the basic facilities already available.

TouchNTalk is also intended to provide enhanced terminal access to arbitrary host facilities using a standard modem to allow data connections to suitably equipped computers at remote locations which would include public access information services. A terminal emulator matched to the system communicates with the host and updates the pseudo-display appropriately. Thus any software that can be operated using a normal terminal window can, in principle, be operated by visually impaired people using *TouchNTalk*.

Problems caused by optimised screen update, which uses control characters and text fragments to minimise the number of characters transmitted when updating the screen, are avoided because the user only accesses the information once it is properly formatted on the pseudo-display. Also, interruptions from clock-time updates and the like do not occur. The user can determine the time by reading the appropriate portion of the pseudo-display pad when it is convenient. This is in contrast to a normal talking terminal, which speaks the characters as they arrive, in whatever order, or—if "silent mode" has been selected—ignores them.

Multiple window operations are possible. The terminal emulator can also equate a page larger than the size of the pseudo-display with the screen of the emulated terminal, and thus emulate a screen of arbitrary size. The window-dependent mechanisms outlined above will provide adequate means of viewing all parts.

What we have not solved are the problems of drawing the user's attention to incoming error messages, unexpected screen updates, and most difficult of all, small changes. With a multiple window view for the sighted user, visual feedback, even in peripheral vision, tends to make even the smallest changes noticeable, without necessarily being distracting. For the visually impaired user, either every change must produce an auditory signal, which could be very annoying (and still

leaves the user with the problem of determining the full extent of the change(s)), or the changes can be ignored, which could be worse. Error messages are a special problem. Since the terminal emulator will pass "bell" characters as appropriate noises, host software can tag important updates by sounding the bell. However, the user still has to find out the full extent of the changes, which could range from a single character to the entire screen, and such a solution requires modification of the host software, or possibly an intelligent agent in *TouchNTalk*.

One reasonable compromise would be to beep once for every continuous string of characters received, update the screen, and keep a secondary pseudo-display showing only the most recent changes, based on differencing the new screen with selected older screens maintained like a history mechanism, and providing the facility for reading them both in and out of the whole screen context using suitable holophrasts. The user could control the rate at which the older screens were updated and thus control the amount and currency of the change information displayed on the secondary display. The idea could obviously be elaborated to make finding changes of any reasonable age fairly straightforward.

Soft functions could be provided: to enable and disable the change beeps (which would be different from the bell character—both could be customisable); to switch between primary and secondary pseudo-displays; and to control the context in which changes were placed. When reading changes in context, speech resources of intonation, rhythm, voice quality and the like could be used to distinguish the context from the changed material. Coupled with suitable bell characters from the host, the user would be in a strong position to judge the likely importance of changes (given knowledge of different current activities), and to find out exactly what had changed, quickly and conveniently. This our current line of development.

A better solution to the problem will depend on either considerable applications-specific intelligence in the workstation, or specially written software in the host that makes use of extended functions (yet to be defined) in the workstation. Other secondary pseudo-displays could be used to provide ready access to other types of information (such as local system information), within reason, and within the ability of the user to cope with all the entities. We have yet to tackle the auditory equivalent of full window systems, although Cohen (1993) provides an exciting start in this area, but not specifically directed at the visually impaired.

6.9 Evaluation

The basic concepts and methods of the *TouchNTalk* system were evaluated by controlled experiment (Hill & Grieb 1988). Blind users, and blindfolded normal subjects performed editing tasks of a direct manipulation nature about 50% faster when using *TouchNTalk* than when using a conventional talking terminal that lacked the multimodal enhancements but that was specially adapted in other respects to the required task characteristics. Blind subjects were also enthusiastic about *TouchNTalk* and correctly believed they performed better using it.

Interestingly, the blindfolded normal subjects subjectively preferred the key

based conventional talking terminal, and they thought they performed better when using that, despite the objective evidence to the contrary. We can only assume that, as they were all frequent keyboard users, the apparently paradoxical preference for the key device somehow reflected their great familiarity with keyboards, compared to special devices with textured surfaces, which they had almost certainly never encountered before. Based on our own (unpublished) experience in other experiments involving blindfolds, we feel that this effect may have been enhanced by the stress of wearing a blindfold. Blind subjects, on the other hand, are used to feeling and exploring without seeing, as part of their everyday activities, including (for some) "reading" Braille books. It seems that the pad device format has potential for blind computer users in the sense that not only does it work, but it is also immediately acceptable and comfortable.

The discrepancy between subjective opinion and reality illustrates a cautionary note in further experiments. Whilst we may assume that performance may be measured in similar tasks using sighted subjects wearing blindfolds as a basis for design, the subjective aspect of users can only safely be investigated using blind subjects. Subjective preference does not necessarily follow conditions for optimum performance. In any case, frequent validation of results obtained using normally sighted subjects must be undertaken using blind subjects, especially if the aspects being investigated, or the format of the experiment, were to differ significantly from our study.

It should also be noted that the context for blind subjects in answering our questionnaires was quite different from that of the normally sighted subjects. This, we speculate, is part of the explanation for the differences in response. All of the blind subjects had used some form of talking terminal before, and, for them, this experiment presented a real situation under real conditions. For the sighted it was more like some strange sort of game. Thus the blind subjects were far more concerned with the usability of the devices. Given their background, the normally sighted subjects, although very familiar with keys, had probably had little or no experience with pointing devices, and certainly had none using such a device without the benefit of sight, as already noted. Although we specifically asked subjects about their experience with talking terminals (in the pretest questionnaire), we overlooked the almost equally relevant question about experience with pointing devices.

7. Hi Fi Mike: animating speech and gesture: multimodal communication using speech, facial expression, and body language

7.1 Introduction

It is well known that the perception of speech is influenced by visual cues as well as auditory cues. Pairing speech sounds with corresponding lip movements

makes them more intelligible (Sumby & Pollack 1954; Benoit in press; Mohamadi & Benoit 1992). The effect is strong enough that the same speech sound may be perceived differently when paired with a different set of lip movements (McGurk & MacDonald 1976). This is why it is worth presenting a visual depiction of a face, along with speech, when using speech output from computers. Such research is also required as a basis for automating the process of animating speaking characters for entertainment purposes. The work provides an example of multimodal communication that is quite different in character and problems from *TouchNTalk*. Amongst other problems, it requires managing precise synchronisation, not just minimising response time.

7.2 Speech and lip synch in computer animation

At the University of Calgary, as a joint project with the *Graphicsland* group, we have been working on a method for automatic speech animation. By adding a few extra parameters to our existing program for text-to-speech synthesis, it is possible to control the lip and jaw movements of a computerised face model to produce suitable articulatory movements perfectly synchronised to the synthetic speech produced (Hill, Pearce & Wyvill 1989; Wyvill & Hill 1989; Wang 1993). The original motivation for the project came from a need to automate the tedious manual process of lip synch in the production of computer animated movies. Traditional methods either ignore the problem (so that mouth movements of characters bear little relationship to their speech) or they photograph a real actor whilst recording a sound track, and then use *rotoscoping* to transfer the mouth movements to the cartoon characters frame by frame as part of the drawing process. It is said that it takes all the fun out of animation.

Some work has been done by others to try and automate the rotoscoping technique directly, using speech recognition techniques to identify the key frames needed for the speech animation, but without a great deal of success in terms of reducing the labour and improving the quality.

The problem with our method is that most synthetic speech quality is unnatural. Although we are working on the problem (Hill, Manzara & Taube-Schock accepted), for purposes of computer animated cartoons there is a simple solution. After the film has been generated, a real voice may trivially be dubbed over the existing synthetic speech soundtrack. We have used this strategy successfully.

7.3 Multimedia speech output

Animated speaking faces and characters provide interesting possibilities for CHI, and have provided the focus for Christian Benoit's presentation at this workshop. The relatively poor quality of the synthetic speech is an obvious potential problem because, unlike the film animation case, dubbing real speech is not possible. Experience has shown (informally) that synthetic speech presents little if any problem in an interactive situation and may even be preferred when interacting with a machine (Witten & Madams 1977; and Witten 1982, p.7). There

is a reasonable expectation that the combination of interactive use and added visual cues may provide a very attractive, robust form of speech output channel for computer use, including specialised applications such as speech therapy. The system has not been tested for interactive use so far because we lack equipment capable of generating and rendering facial images in real-time, although we have experimented with wireframe face models. Like many problems in CHI, this is a problem of technology that we expect to disappear in the foreseeable future.

We do plan to perform experiments to determine the effect on intelligibility of adding the face in an interactive context. As a precursor to this, we are developing a new synthesis system based on an articulatory model to achieve greatly improved naturalness which will also provide articulatory parameters for facial synthesis directly, including parameters expressing tongue position.

Speech animation, as Parke (1982) stated, requires more than stylised lip and jaw movements to be convincing. The appearance of the tongue is also important, and other aspects of facial expression undoubtedly play a role similar to that played by rhythm and intonation (prosody) in the auditory domain. This is one reason why many professionals involved in helping deaf people to use visual cues, in understanding a speaker, prefer the term "speechreading" to "lipreading".

Jeffers and Barley (1971, p.4) amplifying Nitchie's definition of lipreading, define speechreading as the art of understanding a speaker's thought by watching the movements of his or her mouth and facial expression, noting that the process is not exact, and that a speechreader must fill in a great deal of missing information using experience, context, general knowledge and the like, since many sounds and even words (called *homophenes* [sic]) are indistinguishable on the basis of visual information alone. At the same time, they agree that the auditory signal is more comprehensible when accompanied by the sight of the speaker's face and note the importance of body movements (gestures). Missing such information degrades interaction when speaking to people over the telephone or across a shower curtain, and explains why people like to view the speaker at public lectures.

Computer-animation of speech ultimately demands that attention be paid to the non-verbal channels inherent in movements of the entire body, as well as the facial expression.

7.4 Gesture and speech interaction; relation to prosody (and, briefly, another "m" word)

Condon, after a ten year study based on hours of videotapes, characterised all body movements as controlled by muscle action units (MAUs) (Condon 1971; Condon & Ogston1974). Condon's work predates that of Ekman and Friesen (1975; 1977) which led to the FACS system for coding facial expressions. An important unexpected finding was that the beginnings and endings of activity in the MAUs, besides exhibiting varying group synchronisation amongst themselves for body movements, *were also synchronised with major speech events* (phonetically important markers resulting from articulatory processes). This was true not only for

the *speaker's* movements, but also for those of *listeners* movements, and even when the listeners were newly born infants, apparently unable to speak or understand. This finding suggests that, quite apart from the obvious need for lip synchronisation, a mechanism for correlating the body movements of *all those depicted in an animation* with the any concurrent speech is of fundamental importance to realistic communication. Artistic license may even require an exaggeration or caricature of the effect. That cartoonists have long been aware of this need is obvious if one watches any high quality cartoon (for example, Hubley, Hubley & Trudeau 1983). As noted body motion (including facial expression changes) probably plays a visual role akin to that played by prosody in the auditory domain.

The problem of categorising and synthesising body language is another open research question. We don't even know how to control prosody in an appropriate manner for speech synthesised from text. Such control requires solution of the AI-hard problem of language understanding as well as a better characterisation of the relation of intent and meaning to rhythm and intonation. The topic generated considerable interest during discussions at the workshop and Marcel Tatham was much more optimistic than I was about the current state of knowledge. Appropriate rhythm and intonation requires (a) that the text being spoken be understood; (b) that we know what are the relevant features of intonation and rhythm; and (c) that we know how to vary them appropriately for given meanings, situations and emotional effects; and (d) that we can accurately place them in synchrony with arbitrary speech. These are yet more open research questions and form long-term goals for our work.

The appropriate categorisation and control of facial expression and other body movements is even more problematical. In my demonstration video, the gestures and body movements of the character Hi Fi Mike were produced by artistic intuition and implemented manually. However, it is clear that correct automation of such gestures and body movements is just as important and desirable as the automation of lip-synchronised speech with correct rhythm and intonation.

7.5 Facial expression and gesture as a communication modality apart from speech

Facial expression and gesture may also be used in a different modality from speech and body language. For example, Chernoff's work on the presentation of multivariate data using components of facial expression (Chernoff 1973) has been followed up by others (De Soete 1987; Levine 1990). The idea is that judicious choice of the way certain components are drawn, according to the values in each dimension for a point in the space concerned, allows a user to compare the points much more easily than using other representations, because faces are familiar. The example given in Levine (1990), which was concocted by the editors rather than the author, provides a convincing comparison of the difference in heart attack risk factors between a large number of US cities. Figure 4 shows a couple of example faces from Levine. The disposition of heart attack risk factors to facial features has

been so judiciously done that the editors felt compelled to issue a caution with the figure (which showed faces representing over 40 US cities):

> *It is not suggested that the facial expressions reflect the mood of the corresponding cities, but perceived similarities and differences in the faces may offer clues to the nature of life in the cities. (The Editors, Levine 1990)*

It seems entirely reasonable to suppose that the whole body could be used in a similar way, if the perceptually relevant aspects of body position and gesture could be characterised and manipulated. It is probable that certain data is more suitably presented this way than other data and perhaps other images could be used as the substrate. However, Chernoff based his approach on the fact that humans have a particular sensitivity to facial features.

New York, NY   Los Angeles, CA

Two examples of multidimensional data on the pace of life and incidence of heart disease using Chernoff faces.

Walking speed = angle of eyebrows
Talking speed = width of mouth
Frequency of watch wearing = height of eyes
Speed of bank transactions = diameter of pupils
Death rate from heart disease = curvature of mouth

(After Levine 1990)
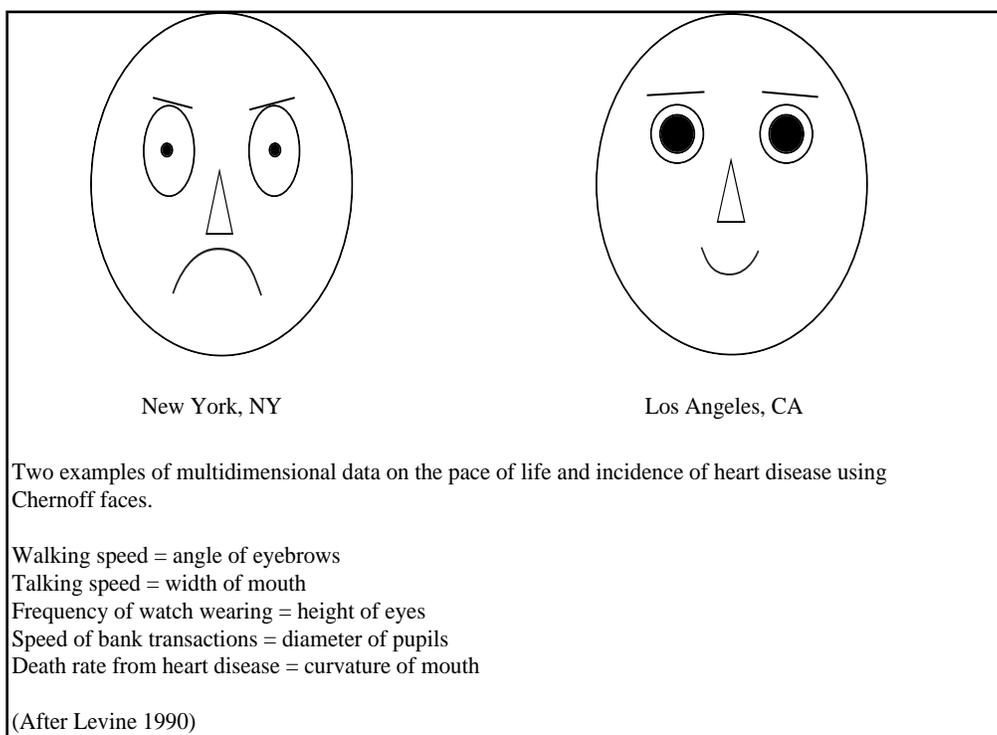
Figure 4: Chernoff faces showing the pace of life and heart disease

8. Comments on the multi-modal design problem

8.1 *TouchNTalk*

It is worth extracting some comments on the multimodal dialogue structure design problem from our experience with two quite different but relatively complex essays in this area—*TouchNTalk* and the animation of speaking characters,

particularly *Hi Fi Mike*.

A question I am frequently asked, in connection with the *TouchNTalk* system, is how did I design it. It is assumed that somehow the design process was regular and methodical and reflects methods that could be applied to other systems. I am afraid that, in general, this is not so, although the design guidelines I have used in teaching CHI for a number of years provided a framework (Hill 1987). The approach to design described by Rasmussen is much closer to reality.

The driving principle was a strong conviction about the *purpose* of the system which was to provide channels of communication, accessible to a blind user, that could be substituted for visual cues; and to ensure that other non-visual cues that even a sighted user would find useful would be preserved in the computer mediated environment for document access that I wished to create. This led to a study of how people access documents, and drew informally on the research carried out by the *Architecture Machine Group* (*AMG*) at MIT as well as explicit research by four other workers (Witten & Bramwell 1985; Benest & Jones 1982). As many cues as could be dreamed up were to be made available to the user, converting modality if necessary.

The central idea was to convert the visual form of a screen, normally accessed using the eyes as a roving cursor, into a different spatial representation that could be viewed by direct manipulation, using a touch surface for access, and synthetic speech output, tactile sensation, and proprioception for feedback, to close the control loop. I felt it was important to try and build in the same redundancy for the blind user as normally exists for the sighted user and, inspired by the *AMG* and the four workers just mentioned, I was acutely aware that conventional document access on computers was cue-deficient even for sighted users.

Apart from text entry, I thought it highly desirable to make access to all facilities uniform in terms of access and control. Text entry I felt could be handled by a normal keyboard if necessary, or—better—using a Maltron one-handed keyboard. The choice of one or two handed keyboard related to a desire to allow the user to keep track of things on the touch surface with one hand whilst entering text with the other. This is not such a major point as it seems, because a blind person using the system tends to use two hands on the touch surface to help keep track of things anyway. Thus even a one handed keyboard tends to disrupt tactile continuity. The ultimate goal is to use speech recognition input, but that opens a whole new can of research worms as far as managing things is concerned. Even after the recognition problems have been dealt with, there are real problems constructing suitable dialogues, especially for continuous speech recognition.

It was clearly important to make *document structure* directly available. Since a large Braille array was out of the question [5], the structure would have to be represented explicitly and accessed using the same mechanism as everything else, but I wanted to relieve the user of the need to track lines of text, if possible. This led

5. Even if a large Braille array were possible, and there have been recent reports of progress in this area, not all potential users can work with Braille.

directly to the holophrast idea.

In checking existing talking terminals we found that one serious problem was the number of functions that had to be controlled. A conventional talking terminal uses keystrokes on a standard keyboard to access everything.   Designers soon run out of keys. As a result, extreme key combinations are used to provide the necessary controls. Such chording became so extreme on one system we looked at that it was physically impossible to span some key combinations.

I did not wish to compromise our design by simply constructing a special function keypad on the touch surface. In any case, we could not afford the pad real-estate, despite our reduced need for special functions.

I decided to use natural gestures wherever possible, although some "key" selections are still used in the function column (it may be possible to replace these); thus spell mode can be signalled by speed of movement rather than by specific selection, even though a "button" is provided. At present, the button allows spell mode to be locked in.

Access to cursors was modelled somewhat on the *emacs* text editor in that the mark was provided to allow a region to be designated, or a previous location to be remembered. *Finding* the cursor drew on my experience as a physicist and pilot. In both these occupations, I learned the value of nulling out tones to zero in on a desired value, position or track. The extension of the idea to indicate book size and the "open here" location was natural and consistent.

Access position (the *user cursor* or *point of regard*) on the pseudo-display needed to be sensed by touch and proprioception, independent of the content (which changes). Textured squares, grooves, and other aids provided tactile cues, whilst kinaesthetic/proprioceptive senses gave a fair idea of location and relative distance. The use of the stylus, instead of straight touch, was serendipitous. We were unable to get a touch tablet with suitable resolution when we started the work, so a *BitPad 1* was used as a compromise, with a specially adapted stylus attached to the finger. Users spontaneously took to using the stylus directly. They were still able to use tactile cues and proprioception, but they obtained a more precise idea of the access position.

Finally, the organisation of the system was important insofar as it made development easy. The move to the *NeXT* was very helpful in this respect because it provided an Object Oriented development environment with excellent support for building UIs. However, organising the knowledge sources into independent experts communicating through something like a blackboard system—akin to Hewitt's actor-based system (Hewitt 1977), was an important aid to development and a model for incorporating expertise into CHI.

The work on *TouchNTalk* from the perspective of the intended blind users can be characterised as Give us the tools (in a suitable form) and we can do the job (of accessing documents). I suggest that this encapsulates the over-riding responsibility of the CHI designer in any system. What I tried to do in my design was to provide tools equivalent to all those modes used by normally sighted readers in real

document reading situations (as opposed to the impoverished situation encountered on most computer systems). I also tried to incorporate the UI design principles I have been teaching to my students for may years (Hill 1987). Finally, I tried to impose an organisation that allowed diverse forms of expertise to be managed without compromising ease of development.

One last comment is worth making. It would have been difficult to make progress on the *TouchNTalk* project without an evaluation phase. Evaluation is yet another key problem in CHI, and may prove especially difficult in Phase 3. Without evaluation, not only is it likely important errors or features of any system will be missed, but there will be no basis for believing that the system offers any advantage to potential users. Regrettably, too many systems end up being evaluated by paying customers, which is not good for anyone.

8.2 Hi Fi Mike: animated characters

The goals and difficulties of character animation were of a different kind to those for *TouchNTalk*. They were concerned with the extremes of Rasmussen's abstraction hierarchy, ranging from difficulties with managing the physical particulars, to difficulties in managing intentional relationships.

The most practical difficulty was our inability to render facial images in real time. Since switching to an articulatory model for speech synthesis, we are afflicted with a similar problem in generating speech on the original 25MHz *Digital Signal Processor (DSP)* in the *NeXT*, and are limited to male voices. This will cease to be a problem with the port to faster hardware, currently in progress, but does raise a general problem in sophisticated CHI—sophistication takes time, perhaps more time than is available. I am reminded of the Doonesbury cartoon seen a few years ago which showed a user choosing between response-time/memory-usage and the degree of user friendliness.

This problem exacerbated the problem of achieving synchronisation between sound and visual media (speech and facial expression modes) and it has not yet been properly solved even when going to film or video. Though we have a procedure that works very well [6], it requires a small amount of manual intervention.   Once the real-time generation problems are solved, synchronisation problems will remain because many operating systems do not provide suitable real-time control primitives—not even the *MACH* operating system used as the kernel for *Unix* on the *NeXT* provides them, which came as something of a shock to me. It would be hard to move to a more primitive operating system in order to gain better control of

6. The procedure involves laying the synthetic sound track onto black video tape (video tape that has been recorded, but with black input, so there are frames without content). The number of frames corresponding to the soundtrack are then counted, and this number is used to determine the video frame generation rate, and the video frames are laid, one at a time, onto the black video. This can cause problems when animation effects that are not related to speech are produced using different procedures, but the problems can, in principle, be solved fairly easily. We have not done so because the real solution is the concurrent generation of both sound and video—which raises the question of synchronising primitives built into the operating system.

real-time events, and this problem will have to be addressed.

Another problem with the facial animation was the tenuous relationship between the parameters available for face control and the facial expression characteristics that one could measure. Mappings were devised on a pretty *ad hoc* basis. Carol Wang's work was designed to resolve some of these problems, and her results for basic animation were superb (Wang 1993). However, the dynamics of the face model were problematical when it came to speaking, since interpolating from one expression to another could produce anomalies. The problems were less severe than when using Water's face model, but work remains to be done. Parke's model mouth collapsed when we tried puckering it (as required for the lip-rounded vowel "oo" (/u/ in the *IPA* notation). Thus our face models (sources of knowledge about how faces look and move) are still deficient. This is entirely in keeping with my assertion that the problem of *models* (i.e. sources of knowledge) will dominate Phase 3 of CHI, as discussed in detail above. This dominance may make CHI almost indistinguishable from some schools of AI, especially as the other major problem, outlined in the presentation above, is that of natural language understanding and the multimodal communication of meaning.

The more serious and fundamental problems associated with the relationship between rhythm, intonation, body movements, and understanding have already been discussed (Section 7.4).

9. Acknowledgements

10. References

ALEXANDER, C. (1964) *Notes on the Synthesis of Form.* Harvard U. Press: Cambridge, Massachusetts

BENEST, I.D. & JONES, G. (1982) Computer emulation of books. *Int. Conf. on Man-Machine Systems* (IEE Conf. Publication 212), UMIST, Manchester, UK, 6-9 July, 267-271, London: IEE

BENOIT, C. (in press) The intrinsic bimodality of speech communication and the synthesis of talking faces. *2nd. Venaco Workshop on the Structure of Multimodal Dialogue*, Acquafredda di Maratea, Italy, Sept. 16-20 (to appear in this volume)

BOLT, R. (1980) "Put that there": voice and gesture at the graphics interface. *Proc. SIGGRAPH 80 Conference* (Computer Graphics 14 (3)), Seattle, July 14-18, 262-270

BOLT, R. (1982) Eyes at the interface. *Proc. Human Factors in Computer Systems Conference*, March 15-17, Gaithersburg, MD: Nat. Bureau of Standards

BOLT, R.A. (1984) *The Human Interface: where people and computers meet.* Belmont, CA: Lifetime Learning Publications (Division of Wadsworth), 113pp, ISBN 0-534-03380-6-Cloth

BRAND, S. (1987) *The Media Lab: inventing the future at MIT.* Penguin Books: New York, London, Victoria, Markham, Aukland 285pp

BROWN, J.S., BURTON, R.R. & BELL, A.G. (1975) SOPHIE: a step towards creating a reactive learning environment. *Int. J. Man-Machine Studies.* 7 (5), 175-218, September

CARROLL, J.M. & THOMAS, J.C. (1982) Metaphor and the cognitive representation of computer systems. *IEEE Trans. on Systems, Man & Cybernetics* SMC-12 (2), 107-116, March/April

CHERNOFF, H. (1973) The use of faces to represent points in a k-dimensional space graphically. *J. American Statistical Assoc.* 68, 361-368

COHEN, M. (1993) Throwing, pitching and catching sound: audio windowing models and modes. *Int. J. Man-Machine Studies* **39** (2), 269-304

CONDON, W. (1974) Speech makes babies move. *New Scientist,* **6** June 1974

CONDON, W. & OGSTON, W.D. (1971) Speech and body motion synchrony of the speaker-hearer. In: Horton, O.L. & Jenkins, J.J. (eds.) *The Perception of Language.* Merril: Columbus, Ohio 150-184

DE BONO, E. (1979) *Future Positive.* Maurice Temple Smith: London

DENNET, D.C. (1971) Intentional systems. *J. Phil.* **LXVIII**, Feb 25

DE SOETE, G. (1987) A perceptual study of the Flury-Riedwyl faces for graphically displaying multivariate data. *Int. J. Man-Machine Studies* **25** (5), 549-555

EKMAN, P. & FRIESEN, W. (1975) *Unmasking the human face.* Consulting Psychologist Press: Palo Alto, California

EKMAN, P. & FRIESEN, W. (1977) *Manual for the facial action coding system.* Consulting Psychologist Press: Palo Alto, California

FOLEY, J.D., WALLACE, V. & CHAN, P. (1984) The human factors of computer graphics interaction techniques. *IEEE Computer Graphics and Applications* **4** (11), 13-48, November

GAINES, B.R. & SHAW, M.L.G. (1986) A learning model for forecasting the future of information technology. *Future Computing Systems* **1** (1), 31-69

GAINES, B.R. (1990) From information to knowledge technology. *Future Computing Systems* **2** (4), 377-407

GENTNER, D. & STEVENS, A.L. (eds.) (1983) *Mental Models.* Hillsdale, NJ: Erlbaum

HALASZ, F.G. & MORAN, T.P. (1983) Mental models and problem solving in using a calculator. *Human Factors in Computing Systems: Proc. SIGCHI 83*, Boston Dec 12-15, 212-216, Baltimore: ACM

HALL, E.T. (1981) *Beyond Culture.* Anchor Press/Doubleday, Garden City, New York, 298 pp

HANSEN, W.J. (1971) User engineering principles for interactive system. *Fall Joint Computer Conference, AFIPS Conference Proceedings* **39**, 523-532, New York: American Federation for Information Processing.

HEWITT, C. (1977) Control structure as patterns of passing messages. *Artificial Intelligence* **8**, (2) 323-363

HILL, D.R. (1987) Interacting with future computers. *Future Computing Systems* **2** (1) 83-124

HILL, D.R. & GRIEB, C. (1988) Substitution for a restricted channel in multimodal computer-human dialogue. *IEEE Trans. on Systems, Man & Cybernetics* **18** (2), 285-304, March/April

HILL, D.R., SCHOCK, C-R & MANZARA, L. (1992) Unrestricted text-to-speech revisited: rhythm and intonation. *Proc. 2nd. Int. Conf. on Spoken Language Processing*, Banff, Alberta, Canada, October 12–16, 1219-1222

HILL, D.R., MANZARA, L. & TAUBE-SCHOCK, C-R. (accepted) Some problems in applying traditional phonetic analysis to speech-synthesis-by-rules. To be presented at the *Int. Cong. of Phonetic Sciences 95*, Stockholm, Sweden, August.

HILL, D.R., PEARCE, A. & WYVILL, B.L.M. (1989) Animating speech: an automated approach using speech synthesised by rules. *The Visual Computer* **3**, 277-289

HUBLEY, J., HUBLEY, F. & TRUDEAU, G. (1983) *A Doonesbury Special.* (Animated cartoon movie). Pacific Arts Video Records: Carmel, California, PAVR-537, 30 mins.

JEFFERS, J. & BARLEY, M. (1971) *Speechreading (Lipreading).* Charles C. Thomas: Springfield, Illinois

KAY, A. (1987) *Doing with Images Makes Symbols.* University Video Communications (sponsored by Apple Computer Inc.) 97 minutes, October

KIERAS, D. & POLSON, P.G. (1984) An approach to the formal analysis of user complexity. *Int. J. Man-Machine Studies* **22** (4), 365-394, April

LEVINE, R.V. (1990) The pace of life. *American Scientist* **78** (5), 450-459

LINDGREN, N. (1966)   Human factors in engineering, parts I & II. *IEEE Spectrum,* **3** (3), 132-139, March; 3 (4), 62-72, April

LUHMAN, N. (1979) *Trust and Power.* Wiley: Chichester

MANZARA, L. & HILL, D.R. (1992) DEGAS: A system for rule-based diphone synthesis. *Proc. 2nd. Int. Conf. on Spoken Language Processing,* Banff, Alberta, Canada, October 12-16, 117120

McGURK, H. & MacDONALD, J. (1976) Hearing lips and seeing voices. *Nature* **264**, 746-748

MOHAMADI, T. & BENOIT, C. (1992) Le gain des lèvres: intelligibilitié auditive et visuelle de la parole bruitée en Français. *Proc. 19° Journées d'Étude sur la Parole,* GCP de la Société Française d'Acoustique, Brussels, Belgium, May

NBC/ACM (1982) *Human Factors in Computer Systems.* Proc. of Conference held at the National Bureau of Standards, Gaithersburg, Maryland, March 15-17

NORMAN, D.L. (1984) Stages and levels in human-machine interaction. *Int. J. Man-Machine Studies* **21** (4), 365-376, October

NPL (1959) *Mechanisation of Thought Processes.* Proceedings of Symposium at the National Physical Laboratory, Teddington, Middlesex, UK, 24-27 Nov 1958, 2 Vols. 980 pp.

PARKE, F.I. (1982) Parameterized models for facial animation. *IEEE Computer Graphics and Applications* **2** (9): 61-68

PFAFF, G.E. (1985) User Interface Management Systems. Berlin: Springer-Verlag, 224pp

POPPER, K. (1963) *Conjectures and refutations: the growth of scientific knowledge.* Routledge & Kegan Paul: London, 431pp

PFAFF, G.E. (1985) *User Interface Management Systems.* Berlin: Springer-Verlag, 224pp

RASMUSSEN, J. (1983) Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man and Cybernetics* **SMC-13** (3), 257-266, May/June

RICH, E. (1983) Users are individual, individualizing user models. *Int. J. Man-Machine Studies* **18** (3), 199-214, March

RISSLAND, E.L. (1984) Ingredients of intelligent user interfaces. *Int. J. Man-Machine Studies* **21** (4), 377-388, November

SEBEOK, T. A. (1986) The evolution of communication and the origin of language. *Distinguished Lecture at the University of Calgary*, May 12th.

SHANNON, C. (1948) A mathematical theory of communication. *Bell Systems Technical Journal,* **27** 379-423; 623-656 July & October

SMITH, D.C. (1975) *Pygmalion: a creative programming environment.* PhD Thesis, Stanford University, June (available as NTIS Report AD-A016 811, Washington: Nat. Tech. Inf. Service)

SUMBY, W.H. & POLLACK, I. (1954) Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Amer.* **26** (2), 212-215

SUTHERLAND, I. E. (1963) Sketchpad, a man-machine communication system. *AFIPS Spring Joint Computer Conference,* Detroit, Michigan, May 21-23

TAUBE-SCHOCK, C-R. (1993) *Synthesizing Intonation for Computer Speech Output.* M.Sc. thesis, Dept. of Computer Science, University of Calgary

TAYLOR, M.M. (1988) Layered protocols for computer-human dialogue I: principles. *Int. J. Man-Machine Studies. 28* (2 & 3), 175-218, February/March

TAYLOR, M.M. (1988) Layered protocols for computer-human dialogue II: some practical issues. *Int. J. Man-Machine Studies.* 28 (2 & 3), 219-258, February/March

THOMAS, J.C. (1978) A design-interpretation analysis of natural English with applications to man-computer interaction. *Int. J. Man-Machine Studies* **10** (6), 651-668, November

THOMAS, J.C. & CARROLL, J.M. (1981) Human factors in communication. *IBM Systems Journal* **20** (2), 237-263

WANG, C. (1993) *Langwidere: a Hierarchical Spline Based Facial Animation System with Simulated Muscles.* M.Sc. thesis, Dept. of Computer Science, University of Calgary

WASON, P.C. (1971) Problem solving and reasoning. *British Medical Bulletin* **27** (4), 206-210

WILLIAMS, M.D. (1984) What makes RABBIT run. *Int. J. Man-Machine Studies* **21** (6), 333-352, October

WITTEN, I.H. (1982) *Principles of Computer Speech.* Academic Press: London, 286pp

WITTEN, I.H. & BRAMWELL, B, (1985) A system for interactive viewing of structured documents. *Comm. ACM.* **28** (3), 280-288, March

WITTEN, I.H. & MADAMS, P.H.C. (1977) The Telephone Enquiry Service: a man-machine system using synthetic speech. *Int. J. Man-Machine Studies* **9** (4), 449-464, July

WYVILL, B.L.M. & HILL, D.R. (1990) Expression control using synthetic speech. *SIGGRAPH '90 Tutorial Notes* **#26**, 186-212