PsychNology Journal, 2003
Volume 1, Number 3, 229 - 241

The Vista Project*: Broadening Access To Digital TV Electronic Programme Guides

A. Carmichael[•♣], H. Petrie°, F. Hamilton° and J. Freeman*

[•]University of Dundee, °City University, *Goldsmiths College (University of London)

ABSTRACT

VISTA is a multidisciplinary/cross-sectoral project aimed at developing a 'virtual assistant' embodying a speech based interface between digital television viewers and the content and functions of an electronic programme guide (EPG). While it is anticipated that the resulting interface will make EPG access easier for all, the main target groups are visually impaired (VI) and older viewers who experience disproportionate difficulty using currently popular GUI style EPGs.

Despite their great potential for improved usability, speech interfaces are unlikely to prove the 'universal panacea' some anticipate. Rather they raise a host of new human factors issues. For example, current technology disallows a truly 'conversational' interface, thus a structured dialogue is required which raises issues about keeping users 'on-script' (e.g. prompts and other additional support) and providing efficient routes to the information users require. Many of the VI population are elderly and thus also have hearing problems which emphasizes the intelligibility of the synthetic speech output. Qualitative results are presented from iterative evaluations involving (non-VI) elderly users and a wide age range of VI users of a PC based prototype designed to be compatible with digital broadcast technology.

Keywords: *digital television, electronic programme guide, virtual assistant.*

Received 20 October 2003; received in revised form 1 December 2003; accepted 9 December 2003.

1. Introduction

In recent years the UK has witnessed the advent and rapid take up of multi-channel digital television delivered via satellite, cable and digital terrestrial broadcasting technologies. Viewers can now choose programmes from literally hundreds of channels as opposed to the 4 or 5 channels available through analogue terrestrial technologies. In addition, with the convergence of telephone, computing, and broadcast technologies, digital television (and currently at least, the accompanying set-top-boxes) may well become *the* communications focus of the home. If so, they will act as platforms both for TV and enhanced information and entertainment services. For many homes they may become the only means of accessing the internet for e-mail, e-

* Corresponding Auhtor:
A. Carmichael
E-mail: acarmichael@computing.dundee.ac.uk

commerce, e-banking, and e-government. Thus the range of services and tasks supported by the TV is likely to increase substantially in the coming years.

Clearly such convergence provides more opportunities for consumers, in terms of service choice but it also brings the potential for additional complexity. In particular, viewers will need to navigate larger and more complex information structures. With hundreds of channels available, even the relatively simple task of finding out 'what's on now' could prove challenging to some viewers. Similarly, finding out what's on later in the evening, or another day, and remembering what programmes are on and at what time, has the potential to become increasingly difficult. New services also bring new tasks, such as setting reminders, saving programmes to a hard-disc; 'buying' programmes on-demand, and so on. As the range of functions, services and the associated user tasks increases it becomes ever more important that the user interfaces to such services are well-designed to enable all users to gain optimum benefit from these new opportunities.

Currently access to programme listings and services through the on-screen Electronic Programme Guide (EPG) is controlled by a 'traditional' TV remote control (or indeed more than one). The design of these EPGs has generally been influenced by the established approach of windows, icons, menus, and pointers. With multiple services and channel listings available over several days, this generally means 'step-by-step' navigation through a very large matrix of information. Of course, this style of interaction does not cause all users problems. Indeed, the services accessed by users who are comfortable using technology, who perceive clear benefits in being able to access them *and* who have the visual ability and manual dexterity required mean that current systems are evaluated relatively positively by such users (Freeman et al, 2003).

However, given the central role that TV plays in the lives of the vast majority of people, it is important that all sectors of society can benefit fully from digital multi-channel television. Barriers faced by people with, for example, sensory and/or motor impairments exist because of inherent features of graphical user interfaces (GUI). Simply 'optimising' the graphical-approach, particularly if based on the concept of the 'average' user, is unlikely to substantially help visually impaired (VI) or elderly viewers (Carmichael, 2002). Further, the cognitive effort required to navigate through

hierarchical menus (Rama et al, 2001), sets of icons (Rogers, 1986) and other traditional features of the GUI approach is somewhat incongruous in a domestic setting where viewers are more likely to want to sit back and relax. Whilst many users may report that some current EPGs provide “a simple and attractive” interface (Rosengren, 1996), it has been reported that using EPGs can be problematic even for some young, unimpaired users (Daley-Jones, 2002).

As a result of recent advances in speech-technology it has become feasible to develop an alternative means of accessing these EPG services (and potentially, other similar data structures). That is to say, the VISTA Project has developed an interface which accepts spoken input and provides (synthetic) spoken output. The interface is visually represented as an on-screen avatar, which incorporates mouth movements synchronised with the speech output and is also capable of various gestures and facial expressions. The system can also access a ‘real-world’ EPG database to interrogate it for times, channels, titles etc. and can present the user with similar information including programme synopses etc. as shown in Figure 1 below.

Although this alternative approach effectively avoids the well known human factors problems of using a GUI with a ‘traditional’ remote control, it does however raise new ones mainly due to the fundamental differences between the modes of communication involved. The rest of this paper describes some of the main human factors issues that have emerged throughout the project. Some of these issues have been addressed by developments within the project while some will remain to be addressed by future research and development and will be identified as appropriate.

User:	***[Switches on VISTA]***
VHI:	Hello. Do you want to watch something now or later?
User:	***Now.***
VHI:	I can search for a programme title or a channel name or I can just read a list of what's on now?
User:	***Inspector Morse.***
VHI:	Inspector Morse starts at seven thirty pm on Sky One. Do you want to hear more or have I got the wrong programme?
User	***Hear more.***
VHI:	Morse and Lewis investigate a murder in Oxford…
User:	***Watch it.***
VHI:	Please wait a moment while I switch to Sky One.

Figure1: VISTA Avatar and example dialogue excerpt

2. Evaluations

Early versions of the VISTA system have been evaluated with a range of users, consisting of both blind and partially sighted users and (non-VI) elderly users. The objectives of these initial evaluations were to understand the users' overall grasp of the concept and to identify usability issues to be addressed in subsequent iterations. One of the major challenges for the evaluations has been the development of suitable dialogue specifications. That is, most other speech based interfaces (to information structures of comparable size) have been developed to replace a human interface (e.g. flight reservations). In such cases, there is usually a large corpus (recorded and transcribed) of the spoken interactions users have had with a functionally equivalent system, which forms an obvious basis for initial dialogue specifications (Bernsen et al, 1998). With no equivalent forerunner, the VISTA Project had to develop the dialogue specifications 'from the ground up'.

Thus, starting with the very basic tasks of 'finding something to watch now' and 'finding something to watch later', the dialogue structure was developed in light of user trials. The other main factors which shaped the dialogues were the structure and format of the real-world EPG data base the system interrogates and the need for 'repair' dialogues to recover from the inevitable recognition errors.

2.1 Participants

All participants involved in the evaluations were drawn from extant volunteer panels administered by the two Human Factors partners involved in the project. That is, The Centre for Human Computer Interaction Design at City University which focused on VI and blind users; and The Age and Cognitive Performance Research Centre at Manchester University which focused on elderly users. The VI volunteer group included a range of visual limitations and had an average age of 34 years (range; 18-48 years). The elderly volunteer group included a 'normal' range of hearing and visual ability (some wore corrective lenses but none had hearing aids) the average age was 69 years (range; 57-91 years). The findings reported below are based on the contributions of forty volunteers at each centre.

2.2 Procedure

User evaluations were run on an individual basis with each session lasting approximately one hour (occasionally slightly longer for the elderly users). Following a brief explanation of the system and its function, users were encouraged to explore the system on their own and then were asked to perform specific tasks, such as; "Can you find something to watch now" and "Find out when Inspector Morse is on next". Generally, the VI volunteers could follow this protocol, whereas many of the elderly volunteers could not. This latter was mainly due to the various disproportionate difficulties the elderly volunteers experienced with the system (more on this below). Thus, overall, the elderly volunteers required significantly more encouragement, prompting, and occasionally intervention by the experimenter. Although the system was initially set up to record logs of the interactions, this relative lack of task and session structure (exacerbated by a variety of system failures) severely limited the utility of this potentially quantitative data. Therefore, at an early stage in the evaluations the emphasis was placed on more qualitative feedback from users. Following the interaction tasks the volunteers were asked about their experience and were encouraged to give any general and/or specific comments about the interface.

As will be seen below, although VISTA is primarily aimed at both elderly and VI users on the basis of their similarities (e.g. difficulties accessing on-screen text-based information and in using a 'traditional' remote control) it became apparent early in the

evaluations that development of the interface would need to accommodate the various diverse needs of these (and mainstream) user groups. Some of this diversity can be addressed by a suitably flexible system, but some may be better addressed by different 'modes' or entirely separate systems.

3 Findings

3.1 Perceived Need

The initial evaluations showed a marked difference in the perceived need for this system between the (relatively young) VI volunteers and the non-visually impaired older volunteers. For example, many VI volunteers following their evaluation sessions at City University made comments such as "*This would open a whole new world to me*." and "*It's instant access… better than trying to scroll through Ceefax pages*". Such comments indicate that the problems these blind and partially sighted volunteers experienced during the evaluations were effectively out-weighed by the perceived benefits of being able to find out what programmes were showing and when. This supports other anecdotal evidence which suggests that the main barrier to VI people watching TV is not the visual nature of the medium itself, but rather the difficulty of accessing information about the content.

By contrast, the elderly volunteers who evaluated the system at Manchester University did experience relatively more problems using the system (more on this below) but showed that these were far from out-weighed by any perceived benefits. That is, although some of the elderly volunteers showed the common tendency to blame themselves for the systems failings (Levy, 1996), others were somewhat more forthright. For example, one elderly volunteer said, "*It's very hard and she's stupid… it's far quicker to look in the TV Times.*" Alongside indicating a lack of perceived need, this and similar comments also indicate a lack of awareness about various implications of digital television (despite explanations and numerous prompts during evaluation sessions). Given that the current capacity for digital TV in the UK is around 400 channels, paper based listings (such as *TV Times*) are becoming increasingly unfeasible, particularly with regard to informing the viewer of more than just the start time and title of a programme. Despite various attempts to explain the potential benefits of digital TV to such older volunteers it was apparent that some felt that more

channels simply represented "…*more rubbish and repeats*…" and that they were happy to stick with what they are familiar with. Whether this will remain a tenable position following the analogue switch-off seems unlikely. However, it highlights the importance of ensuring easy access for all to information about the content of TV if sections of the population are not to be disenfranchised.

3.2 Vision

It was perhaps obvious at the outset of the project that people with no vision have no need for the visual element of this interface. However, this group represents a fairly small minority of the VI population, and others may derive some benefit from, or simply prefer the inclusion of, the avatar (and other potential variations within the visual element as discussed below). It is likely that for a future 'product' many potential users will consider the trade-off between the potential benefits of the visual element and the associated (monetary) cost. That is, aside from the infrastructure for integration with EPG data structures and basic TV functions, a no-vision version of this system would require effectively only a microphone as the remote control. Whereas, the 'full' version would require some method for presenting the visual element which would likely have some associated extra cost. The project has not specifically addressed such issues but volunteer comments suggest possibilities like picture-in-picture (on-screen) or a small screen on the (PDA like) remote control. Another aspect of a no-vision (or rather, voice-only) version is the possibility of remote interaction via a mobile phone which some 'telly addicts' may find beneficial.

3.3 Hearing

Beyond simply embodying the 'assistant', one of the roles of the avatar was its possibility for helping the many older people who, due to presbyacusis (i.e. age related hearing loss *and* 'slowing' of signal processing), may have some difficulty with the synthetic speech output (Helfer & Wilber 1988). It is known that (particularly older) people can benefit from 'speech reading' when listening to degraded or 'noisy' speech (Sumby and Pollack, 1954), although the benefits are less certain if the face/mouth is not faithfully reproduced, for example because of low frame rate (Williams & Rutledge, 1998). Unfortunately technical compatibility issues have meant that an overall delay in 'lip synch' has only been removed in the most recent version (using 24 'visemes'), for which evaluations are planned.

As mentioned above, the elderly volunteers generally experienced more difficulty using the system than did the younger VI volunteers. Some of this may be due to different levels of motivation (perceived need) during the evaluation sessions, but much qualitative evidence from the evaluations suggest that limited redundancy in the synthetic speech played a major role. One example of this relates to the lip movement delay mentioned above. That is, the quantity and nature of the elderly volunteers' complaints about this strongly suggest that (consciously or not) they were attempting to 'speech read' in order to ameliorate the limited intelligibility of the avatar's speech but were thwarted by the delay.

3.4 Implicit & Explicit Prompts

As can be seen from the dialogue excerpt in Figure 1, the current system has a relatively fixed dialogue structure with the system utterances implicitly prompting the user with the 'commands' they can use. This limitation is mainly due to the necessity for all titles and channel names to be in the system's active recognition 'vocabulary'. Although this is unlikely to be the optimum approach for the future, most volunteers adapted quickly to the convention and generally responded appropriately. However it soon became apparent that many of the older volunteers failed to respond to the implicit prompts, and were often left unsure what to say. This would often lead to out-of-vocabulary utterances and/or filled pauses (e.g. "er"s and "uhm"s) 'contaminating' otherwise in-vocabulary utterances. These recognition problems would in turn lead to further difficulty. This problem has been reduced to a great extent by the inclusion of an 'explicitly prompted' version which can be selected by the user when the system is started up. Thus, the 'default' version is as reported in Figure 1, whereas the prompted version follows the same structure but appends an explicit prompt to each system utterance (e.g. "…Do you want to watch something now or later? ***Please say now or later***."). Following the implementation of this, the elderly volunteers' reactions suggest that many would use it like a 'training mode' whilst some may require it as a permanent part of the system.

3.5 Text Support

The most recent version of the system also includes the possibility of 'text support' such that the speech output is duplicated as on-screen text. This followed

suggestions from elderly volunteers who claimed their difficulty was due less to the intelligibility of the synthetic speech *per se*, but more to their difficulty in *remembering what to say* (once they had decided what they wanted *to do*). Thus the on-screen text can both support the speech output as it appears and act as an memory aid after the transient speech wave has disappeared. Not only does this approach have potential benefits for some older people but also for others such as those with other types of hearing problem (so long as they can speak reasonably well) or those who have English as a second language.

3.6 Synthetic Speech

Partly as a further enhancement of the intelligibility of the speech output and partly as improvement in general acceptability would be the incorporation of suitable prosody/intonation in the synthetic speech. In regard to acceptability many volunteers commented that the speech sounded rather 'robotic' and that they would prefer it to sound more 'natural'. Although mostly general, such comments seemed disproportionately aimed at those instances when the system reads out a programme synopsis. It seems likely that this stems from familiarity with the common practice of many television announcers to reflect the 'mood' of the programme in their intonation. Beyond this, there is good reason to believe that better synthetic prosody would generally improve the acceptability and 'understandability' of synthetic speech. In addition there is the possibility that appropriately 'exaggerated' prosody may be of particular benefit to older people. However, this is beyond the bounds of the current project and whilst there are some possibilities that may be implemented in the near future there is still a very large gap between knowledge of what occurs in human speech and how this translates to the generation of synthetic speech (Hirschberg, 2002).

3.7 Role of the Avatar

As mentioned above, the avatar has had its lip movements connected to the speech output, other than this it has until recently only displayed 'ambient' movement (e.g. slight body 'sway' and quasi-random eye blinks). The most recent version however allows the incorporation of other gestures and facial expressions, which can also be linked to the speech output. Similar to that described for prosody above these elements have the potential to enhance the interaction in a range of ways, from

supporting comprehension of the speech output to making the avatar more personable. For example, comprehension may be improved by appropriately placed 'beat' gestures, which are often used by speakers to mark clause boundaries and the like. Similarly changes in eye gaze can be used to facilitate turn taking in conversations. However, as with prosody, there is a lack of knowledge as to how these could be used to best advantage and what is known strongly suggests that 'inaccurate' implementation can be more damaging than none.

Less directly related to the speech output *per se*, other gestures may improve interactions with this system. One banal example is for the avatar to wave as she greets the user at start up (see Figure 1). Other possibilities include a suitable degree of 'frown' to distinguish 'repair' dialogues ("I'm sorry, I think you said…") from 'normal' ones, and other suitable gestures to indicate system activity which the user may otherwise take as a (frustrating) delay. Examples of this could be a 'thoughtful' gesture (hand-on-chin) during recognition delays, or looking off to the side during a 'search' delay. At the most general level it might be beneficial to include movement and expressions that would give the avatar some degree of 'personality' (an element many volunteers commented that she lacked). The issue of the avatar's 'personality' (including appearance), although well beyond the scope of the present project, is an interesting one as there was virtually no consensus among the preferences expressed by volunteers (those who made negative comments about the avatar were asked to suggest how they would improve it). Such preferences ranged from "a nice teddy bear" to "my grandson". While it may be technically feasible to customise the appearance of the avatar, associating that appearance with an appropriate 'personality' and speaking voice, may prove not to be.

3.8 Future Improvements

In general the issues described so far have emerged in response to the requirements of target users in order to allow them to effectively access the systems functionality. However, other issues have been raised by the requirements of target users for whom basic access has proved less problematic and thereby have offered some insight into the rather more nebulous domain of the 'efficient' use of the system's functionality (including potential functionality). In the main these latter issues were raised by the evaluation sessions involving VI volunteers, although those involving the older volunteers also contributed. At present these issues are

necessarily somewhat abstract and certainly beyond implementation in the present project, but could prove important to future developments in this area, and will be briefly described below.

The overarching issue of this type relates to the dialogue structure. Various factors suggest the need for a more flexible structure. For example, many volunteers who completed tasks involving 'browsing' for information on programmes several days away, commented that it often seemed rather laborious. On the one hand it seemed relatively more acceptable if they were simply 'browsing', while on the other hand it seemed less acceptable if they were effectively 'searching'. That is, 'searching' implies that you already have some criteria and that if you do you should be able to give it to the system 'all-at-once'.

Handling such 'compound' enquiries is possible for current technologies. However, as touched on above these tend to involve relatively restricted 'vocabularies' that are effectively 'fixed' (e.g. airport names). In the domain of EPGs a similar approach would need to be able to distinguish (for example) between an enquiry about programmes showing on Friday night and one about programmes with the title *Friday night*. Development of a suitably flexible dialogue structure will take considerable research effort and the VISTA project has taken initial steps in this direction with colleagues at the University of East Anglia undertaking (albeit necessarily small scale) 'Wizard of Oz' studies in order to indicate the ways people may want to use such an interface without the constraints of real 'speech understanding' technology. In addition to the dialogue structure itself similar effort will likely be required to develop suitable 'repair' dialogues which will require an entirely different approach to those developed for the current system.

Despite the potential difficulties it is apparent that flexibility of this kind will be needed to accommodate the diversity of people's requirements and the diversity of ways they may want to use such an interface. Accommodation of such diversity will be vital if an interface of this sort is going to gain wide acceptance rather than be viewed as effectively 'assistive technology'.

Another issue that is likely to be important for future development is the introduction of 'intelligence', one aspect of which would likely be involved in the development of the

flexible dialogues described above. Other aspects could include the ability to make suggestions for viewing options based on the user's prior behaviour and/or their explicitly stated preferences. Other aspects likely to require some form of 'intelligence' relate to the fact that often televisions do not have a single user. This raises issues such as different sets of preferences (including possibly a 'family' set), and handling potential 'clashes' (e.g. two household members book pay-to-view programmes due to be screened at the same time). However it seems unlikely that even the most sophisticated 'intelligence' will be able to solve the speech based equivalent of fighting over the remote control.

Finally, as the functionality of digital TV increases with interactive programme content and the like, and with the prospect, as mentioned in the introduction, that the TV could very well become an important access point for various internet based services including e-government (or rather, t-government) important decisions will need to be made about which interactions/transactions can be effectively controlled by voice and which will be more suited to other modes of input/output.

4. References

Bernsen O., Dybkjær L. & Dybkjær H. (1998) *Designing interactive speech systems: from first ideas to user testing*, Springer-Verlag London, UK.

Carmichael A. (2002) Talking to your TV: a new perspective on improving EPG usability for older (and younger) viewers. *UsableiTV*, 3, 17-20.

Daly-Jones O. (2002) Navigating your TV: The Usability of Electronic Programme Guides. *UsableiTV*, 3, 2-6.

Freeman J., Lessiter J., Williams A. & Harrison D. (2003) *2002 Easy TV Research Report*. http://www.itc.org.uk/uploads/Easy_TV_2002_Research1.doc

Helfer K. & Wilber L. (1988) Speech understand-ing and aging. *The Journal of the Acoustical Society of America*, 83, 859-893.

Hirschberg J. (2002) Communication and prosody: Functional aspects of prosody *Speech-Communication*, 36, 31-43.

Levy B. (1996) Improving memory in old age through implicit self stereotyping. *The Journal of Personality & Social Psychology*, 71, 1092-1107.

Rama M., de Ridder H. & Bouma H. (2001) Technology generation and age in using layered user interfaces, *Gerontechnology*, 1, 25-40.

Rogers Y. (1986) Evaluating the meaningfulness of icon sets to represent command operations. In Harrison M. & Monk A. (eds.) *People and computers: Designing for usability*. London: Cambridge University Press.

Rosengren J. (1996) Electronic programme guides and service information" *Philips Journal of Research*. 50, 253-265.

Sumby W. & Pollack I. (1954) Visual contribution to speech intelligibility in noise. The Journal of the Acoustical Society of America. 26, 212-215.

Williams J. & Rutledge J. (1998) Frame rate and viseme analysis for multimedia applications to assist speechreading. *Journal of VLSI Signal Processing Systems For Signal Image and Video Technology*. 20, 7-23.

***ACKNOWLEDGEMENT:** The VISTA project is led by the Independent Television Commission (ITC) and is part funded by an EPSRC/ESRC DTI PACCIT LINK grant (L328253047/THBB/C/003/00020). Project consortium; ITC, British Sky Broadcasting, Sensory Inc., Televirtual Ltd., City University, The University of Manchester/University of Dundee, and the University of East Anglia.