*Volume 21 (2002), number 4 pp. 713–722* COMPUTER GRAPHICS *forum*

# A Spoken Dialogue System for Navigation in Non-Immersive Virtual Environments

M.D.J. McNeill, H. Sayers, S. Wilson and P. Mc Kevitt

Faculty of Informatics, University of Ulster, Ulster, Northern Ireland

**Abstract**

*Navigation is the process by which people control their movement in virtual environments and is a core functional requirement for all virtual environment (VE) applications. Users require the ability to move, controlling orientation, direction of movement and speed, in order to achieve a particular goal within a VE. Navigation is rarely the end point in itself (which is typically interaction with the visual representations of data) but applications often place a high demand on navigation skills, which in turn means that a high level of support for navigation is required from the application. On desktop systems navigation in non-immersive systems is usually supported through the usual hardware devices of mouse and keyboard. Previous work by the authors shows that many users experience frustration when trying to perform even simple navigation tasks — users complain about getting lost, becoming disorientated and finding the interface 'difficult to use'. In this paper we report on work in progress in exploiting natural language processing (NLP) technology to support navigation in non-immersive virtual environments. A multi-modal system has been developed which supports a range of high-level (spoken) navigation commands and indications are that spoken dialogue interaction is an effective alternative to mouse and keyboard interaction for many tasks. We conclude that multi-modal interaction, combining technologies such as NLP with mouse and keyboard may offer the most effective interaction with VEs and identify a number of areas where further work is necessary.*

**ACM CSS:** I.3.6 Computer Graphics Methodology and Techniques—*Interaction and Techniques*, I.3.7 Three-Dimensional Graphics and Realism—*Virtual Reality*, I.2.7 Natural Language Processing—*Speech Recognition and Synthesis*

## 1. Introduction

Navigation is the process of moving around an environment, deciding at each step where to go [1]. It is a core functional requirement for virtual environment applications and has been identified as the default behaviour which users return to, for example after carrying out interaction tasks such as the manipulation of an object within the environment [2]. Users require the ability to move, controlling orientation, direction of movement and speed, in order to get to desired positions within a virtual environment (VE) [3]. VE applications often place a high demand on navigation skills [4], which means that a high level of navigational support is required from the interface. VE users have varying objectives when using different applications depending on their context. For instance, users of 3D games will have different requirements from users of a 3D information visualization application or users interacting with others in a collaborative virtual environment (CVE). Users of all VE systems, however, require the ability to navigate through the environment, and to interact with objects in that environment in an efficient and error-free manner. The interface and the input devices used to carry out tasks are important determinants of just how intuitive the process will be. While immersive systems and computer games use specialized hardware devices for interaction with the VE (e.g. datagloves or joysticks) desktop systems

typically rely on a 2D screen interface and general-purpose hardware (mouse, keyboard) for navigation. These devices may not, however, be the most effective for navigation in 3D applications. There has been much research on navigation in VEs dating back some 20 years, suggesting that this is an area that presents many problems. Earlier work by the authors and others reports that users experience a number of frustrations when navigating through VEs:

- Lack of support for control of velocity [3,5]. Getting lost or becoming disoriented in the environment frequently occurs when the speed of navigation is too fast or the direction of movement is not as expected [6]. While faster speeds may reduce time and effort in movement over large distances, slower speeds have been found to be better in attaining precise movements [7]. Navigation speed, therefore, needs to be appropriate for the size of the scene and the tasks which have to be carried out [5,8].
- Problems relating to navigational 'modes', for example 'walking' and 'flying' [7,8]. Even when in walk mode, users can become disoriented if movement is not confined to a level plane [9].
- Lack of identifiable landmarks in the VE — landmarks are distinctive environmental objects which can act as reference points and cues in wayfinding [10].
- Support for automatic navigation ('teleporting') to predefined locations [5]. Although this has been found to be an effective means of moving to specific locations, it can increase a user's sense of disorientation if further indicators of position within the environment, such as a map, are not facilitated [8,9].

Usability is concerned with how easy a system is for the user to understand and use and how efficient that system is [2]. Usability measures how well users can carry out their tasks or meet goals when using an application and therefore affects its overall acceptability. The navigational problems outlined above have been shown to result in user frustration and consequently low usability [2].

Earlier work by the authors highlighted the centrality of the navigation process to VE applications, where evaluation of a number of common interfaces showed a considerable degree of user frustration [9]. Experimental results showed that virtually all participants experienced navigation problems with all interfaces. Problems reported included the inability to control the speed of movement with many users experiencing frequent collisions with objects and disorientation; lack of success in achieving small, precise movements; difficulty with turning and difficulty with maintaining suitable viewing positions and orientations. Included in the recommendations derived from these experiments was an identification of the need for an investigation into the effectiveness of other input modalities.

Much work has been reported on the advantages of multi-modal interfaces to navigation in VEs [11,12], although recently this work has concentrated on using eye-tracking technology to support gaze-directed navigation in conjunction with mouse interaction. There is long-standing evidence to show that a multi-modal system which supports active participation by users is better than one which does not [13].

Speech, arguably the most natural form of human communication, can be used as a mode of interaction between humans and computer systems. It can be used in hands-free situations or as an extra input mechanism, and is used in an increasingly diverse range of applications. Golightly *et al.* [14] identified that speech had much potential to support problem solving tasks and identified the particular category of navigation in VEs as a possible area for its use. To the authors' surprise, little evidence of research into the use of spoken dialogue technology to the general problem of navigation in VEs has been found. Some researchers have investigated natural language interaction for specialized applications (e.g. surgery training [15]); the use of speech to navigate through menu systems has also been reported [16]. Early work was reported [17] on using speech input to create VEs, but this work does not seem to have been extended to navigation. A system for navigation by text-based query in virtual worlds has also been described [11], where the need for environment authors to provide for additional annotation in the scene description in order to efficiently and effectively support such a system was identified.

Our work is intended to investigate the effectiveness of spoken dialogue technology as an interaction style to support efficient navigation in (non-immersive) VEs. We begin with a review of intelligent multimedia systems — systems which combine speech with other interaction modalities.

## 2. Intelligent Multimedia

Intelligent Multimedia, which involves the computer processing and understanding of perceptual input from a number of sources such as speech, text and visual images, and then reacting to it, is complex and involves signal and symbol processing techniques from not just engineering and computer science but also artificial intelligence and cognitive science [13,18,19]. With IntelliMedia systems, people can interact in spoken dialogues with machines, querying about what is being presented and controlling the application. Recent work involves interpreting gestures, body language and eye movements.

Although there has been much success in developing theories, models and systems in the areas of natural language processing (NLP) and vision processing (VP) [20,21] there has been little progress in integrating these two subareas of artificial intelligence (AI). Although in the beginning the general aim of the field was to build integrated language and

vision systems, few actually were, and these two subfields quickly arose. It is not clear why there has not already been much activity in integrating NLP and VP. Is it because of the long-time reductionist trend in science up until the recent emphasis on chaos theory, non-linear systems and emergent behaviour? Or, is it because the people who have tended to work on NLP tend to be in other departments, or of a different ilk, from those who have worked on VP? Dennett [22] says:

> "Surely a major source of the widespread skepticism about 'machine understanding' of natural language is that such systems almost never avail themselves of anything like a visual workspace in which to parse or analyze the input. If they did, the sense that they were actually understanding what they processed would be greatly heightened (whether or not it would still be, as some insist, an illusion). As it is, if a computer says, 'I see what you mean' in response to input, there is a strong temptation to dismiss the assertion as an obvious fraud."

People are able to combine the processing of language and vision/graphics with apparent ease. In particular, people can use words to describe a picture, and can reproduce a picture from a language description. Moreover, people can exhibit this kind of behaviour over a very wide range of input pictures and language descriptions. Although there are theories of how we process vision and language, there are few theories about how such processing is integrated. There have been large debates in Psychology and Philosophy with respect to the degree to which people store knowledge as propositions or pictures [23,24].

There are at least two advantages of linking the processing of natural languages to the processing of visual scenes. First, investigations into the nature of human cognition may benefit. Such investigations are being conducted in the fields of Psychology, Cognitive Science and Philosophy. Computer implementations of integrated VP and NLP can shed light on how people do it. Second, there are advantages for real-world applications such as VE applications. The combination of two powerful technologies promises new applications: automatic production of speech/text from VEs; automatic production of VEs from speech/text; and the automatic interpretation of VEs with speech/text. The theoretical and practical advantages of linking natural language and vision processing have been described by Wahlster [25].

Early work for synthesising simple text from 2D images was conducted [26] where an algorithm capable of labelling edges and corners in images of polyhedra was reported. The labelling scheme obeys a constraint minimization criterion so that only sets of consistent labellings are used. The system can be expected to become 'confused' when presented with an image where two mutually exclusive but self-consistent labellings are possible. This is important because in this respect the program can be regarded as perceiving an illusion such as what humans see in the Necker cube. However, the system seemed to be incapable of any higher-order text descriptions. For example, it did not produce natural language statements such as "There is a cube in the picture."

A number of natural language systems for the description of 2D image sequences have been developed [27,28]. These systems can verbalize the behaviour of human agents in image sequences about football and describe the spatio-temporal properties of the behaviour observed. Retz–Schmidt [29] and Retz–Schmidt and Tetzlaff [30] describe an approach which yields plan hypotheses about intentional entities from spatio-temporal information about agents. The results can be verbalized in natural language. The system called REPLAI-II takes observations from image sequences as input. Moving objects from 2D image sequences have been extracted by a vision system [31] and spatio-temporal entities (spatial relations and events) have been recognized by an event-recognition system. A focussing process selects interesting agents to be concentrated on during a plan-recognition process. Plan recognition provides a basis for intention recognition and plan-failure analysis. Each recognized intentional entity is described in natural language. A system called SOCCER [31,32] verbalizes real-world image sequences of soccer games in natural language and REPLAI-II extends the range of capabilities of SOCCER. Here, NLP is used more for annotation through text generation with less focus on analysis.

MaaS *et al.* [33] describe a system, called Vitra Guide, that generates multimodal route descriptions for computer assisted vehicle navigation. Information is presented in natural language, maps and perspective views. Three classes of spatial relations are described for natural language references: (1) topological relations (e.g. in, near), (2) directional relations (e.g. left, right) and (3) path relations (e.g. along, past). The output for all presentation modes relies on one common 3D model of the domain. Again, Vitra emphasizes annotation through generation of text, rather than analysis, and the vision module considers interrogation of a database of digitized road and city maps rather than vision analysis.

Some of the engineering work in NLP focusses on the exciting idea of incorporating NLP techniques with speech, touchscreen, video and mouse to provide advanced multimedia interfaces [34,35]. Examples of such work are found in the ALFresco system which is a multimedia interface providing information on Italian Frescoes [36,37], the WIP system that provides information on assembling, using, and maintaining physical devices like an espresso machine or a lawnmower [38,39], and a multimedia interface which identifies objects and conveys route plans from a knowledge-based cartographic information system [40].

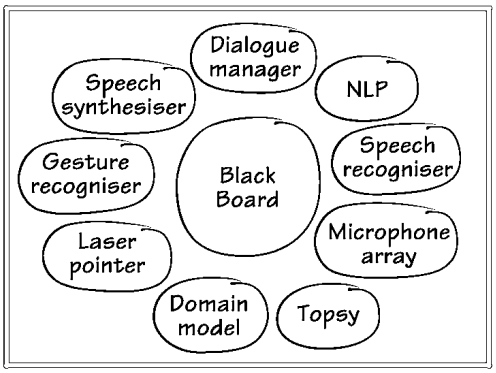


**Figure 1:** *Architecture of CHAMELEON.*

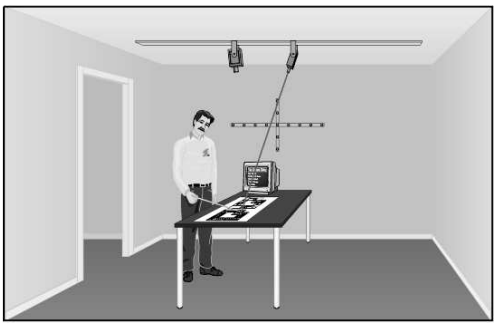

**Figure 2:** *IntelliMedia workbench.*

A general suite of tools in the form of a software and hardware platform called CHAMELEON has been developed (see Figure 1). This can be tailored to conduct IntelliMedia in various application domains [41–46]. CHAMELEON has an open distributed processing architecture and includes ten agent modules: blackboard, dialogue manager, domain model, gesture recognizer, laser system, microphone array, speech recognizer, speech synthesizer, natural language processor, and a distributed Topsy learner. Most of the modules are programmed in C and C++ and are glued together using the DACS communications system. In effect, the blackboard, dialogue manager and DACS form the kernel of CHAMELEON. Modules can communicate with each other and the blackboard which keeps a record of interactions over time via semantic representations in frames. Inputs to CHAMELEON can include synchronized spoken dialogue and images and outputs include synchronized laser pointing and spoken dialogue.

An initial prototype application of CHAMELEON is an *IntelliMedia WorkBench* (see Figure 2) where a user can ask for information about things (e.g. 2D/3D models, pictures, objects, gadgets, people, or whatever) on a physical table. The current domain is a Campus Information System for 2D building plans which provides information about tenants, rooms and routes and can answer questions like "Whose office is this?" and "Show me the route from Paul Mc Kevitt's office to Paul Dalsgaard's office." in real time. Further details are available on `http://www.infm.ulst.ac.uk/~paul/`.

Other work on general IntelliMedia platforms includes *Situated Artificial Communicators* [47], *Communicative Humanoids* [48,49], AESOPWORLD [50,51] and Multimodal Interfaces like INTERACT [52]. Recent moves towards integration have also been reported [13,18,53,54].

## 3. Navigation

Navigating an unfamiliar environment, whether virtual or physical, involves a combination of cognitive processes and motor functions as environmental cues in the environment are evaluated with respect to some overall goal. A number of taxonomies have been reported, including a high-level taxonomy of motor aspects [55] and a task-based taxonomy [8]. Work done suggests that it is feasible to map actions from the real world to actions in the virtual world [10,56]. VEs themselves can be classified in terms of size, density and activity [56]. It is clear that newcomers to an environment rely heavily on landmarks as points of reference [10]. As users gain familiarity with the environment, they acquire route knowledge that allows them to navigate from one point in the environment to another. Route knowledge is acquired and expanded by associating navigational actions and relations to landmarks, such as "turning (action) right (relation) at the Chrysler Building (landmark)". Vinson [10] proposes a number of guidelines for authors of VEs concerning the density, type and uniqueness of landmarks. The process of navigation, when vocalized as in a spoken dialogue system, can perhaps best be thought of in the context of giving directions to a third party. In a sense this is what is implicitly happening in a VE, where the user is giving the system directions (either through the mouse or other modality) to move an avatar (visible, or not) in the VE.

In our system, described below, directional relations are combined with actions and nouns to form commands, which when spoken result in an associated change to the position or orientation of the user in the VE. Example relations include 'left', 'right', 'up' and 'down'. Example actions include 'walk', 'fly' and 'look'. Nouns are used to describe objects in the environment, or landmarks. Landmarks are not only buildings [56], but are paths, edges (e.g. walls, fences), subsections of the environment (e.g. the dockyards or city-centre) and districts (e.g. Capitol Hill, le Quai D'Orsay). Typical spoken commands recognized by our system are:

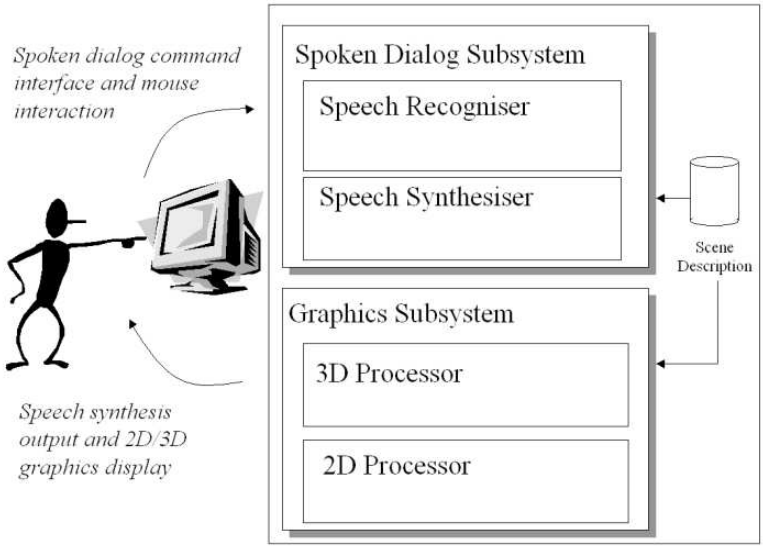


**Figure 3:** *System architecture.*

- "walk to the admin building"
- "look up"
- "turn left"

The issue of navigation in VEs is made more complex as it is possible, and in most cases desirable, to do things which are not possible in the real world. Rapid and non-linear alteration of speed of movement, moving/seeing through walls, jumping (teleporting) from one location to another and viewing a "bird's eye" map are all actions supported in many VE applications. Currently our system supports teleporting (e.g. "jump to") and speed control (e.g. "increase speed"). We note that recent work is exploring a number of new navigational metaphors beyond movement in a linear fashion [8] and expect that this will impact on the direction of our work in the future.

## 4. A Spoken Dialogue System for Navigation in VEs

The system developed consists of five main elements: the spoken dialogue system, incorporating speech recognition dictionaries and speech synthesis components, the 2D graphics interface, the 3D graphics subsystem and the VEs through which the user is navigating. The architecture is shown in Figure 3.

The speech synthesis engine provides audible feedback to the user confirming that the spoken command has been recognized. This is for three reasons:

- it indicates to the user that a particular spoken command had been recognized (or not), while recognized commands are passed to the graphics system for processing. If the system fails to recognize a command it responds with a synthesized voice ("Please speak command again"). Continued failures are met with more informative responses ("Please speak more slowly") and eventually the user is requested to view the help files which list all recognized commands,
- it indicates to the user the particular phrase that the system recognized (not necessarily what was spoken),
- at a subconscious level it reminds the user of the speech interface.

Users can switch the speech synthesis feature off at any time, although in trials so far no-one has opted to do this. However, we suspect this could be irritating after extended use of the system or as user confidence in the system grows.

In common with many desktop interfaces, we have adopted a modal navigation technique. Users can choose between *walk*, *pan*, *look* and *fly*. These modes are common

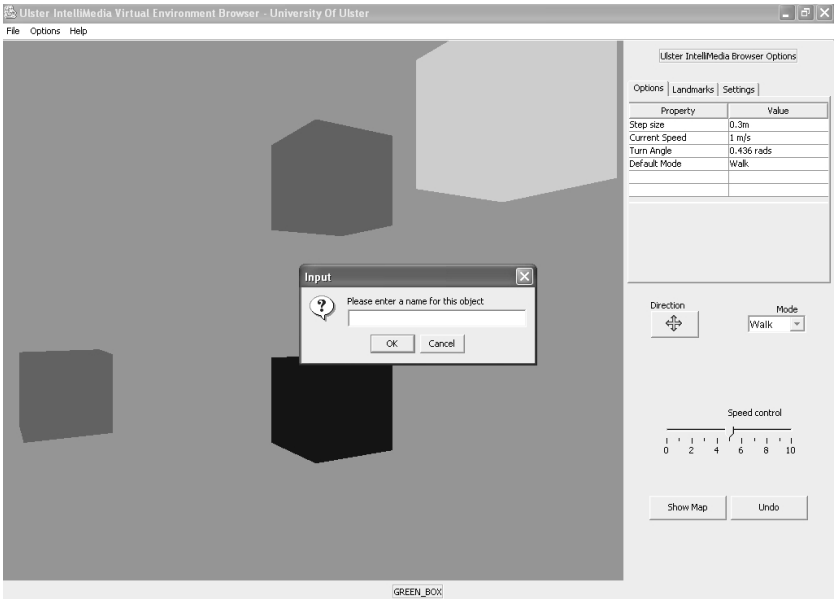


**Figure 4:** *Visible interface.*

in many interfaces — in walk mode, for example, the user is tethered 2m above the ground. Although the user can change modes using the mouse by clicking on the preferred mode in the 2D interface, he or she can also issue a spoken command, e.g. "Change to fly mode". We have chosen to implement a weak modal system, as we believe this enables a more natural form of interaction. When in walk mode, for example, users still have access to look commands. This supports multi-modal interaction styles such as, for example, *rubbernecking* [8], where users can be walking forwards and looking around at the same time. In order to implement this a number of commands are multi-threaded. After a "walk forwards" command has been spoken the system continues walking until a "stop" command is spoken. During this time the user can issue a "look around" command (or "look left", "look up" etc. . . . ) — the effect mimics what many people naturally do in the physical world. When in a particular mode the user can ignore the first part (i.e. the modal element) of the command. For instance, when in walk mode the user can just say "forwards" and the system will respond, thereby reducing the cognitive load on the user. The interface is shown in Figure 4.

A number of simple relations have been implemented: *left*, *right*, *up*, *down*, *forwards* and *backwards*. These can be combined with any of the modes described above so, for example, users can say "walk backwards", "look up" or "pan left".

Where possible our system supports the use of landmark names. The work of van Ballegooij and Eliens [11] details how some scene description languages provide limited support for the naming of objects (landmarks) in the VE. In VRML, for example, it is possible to use DEF-name constructs to give objects an identifier (name). When these identifiers exist in the scene description, our system dynamically builds a dictionary of such names which can then be used to simplify navigation. A small pane at the bottom of the display window shows the name of the landmark as the mouse is passed over it. The user can then use any of the navigational modes to, for example, "walk to the green cube", or "look at the red car". This provides a fast and efficient way of navigating to landmarks of particular interest, greatly reducing both the cognitive and motor loads on users. Further, our system allows users to dynamically rename landmarks. This feature could be of particular benefit in, for example, interactive kiosk applications, where a tourist may wish to rename his hotel 'home' when interacting with a virtual city. Although we could have supported this renaming of landmarks using speech term learning we elected instead for a more traditional route and use mouse and keyboard interaction. Speech term learning incurs the problem of how to correctly spell recognized words (e.g. *kar* instead of *car*). In our system, if the mouse button is pressed when pointing to an object, a dialogue box prompts the user for a new name for the landmark, which the user can then type. These new names are then shown on

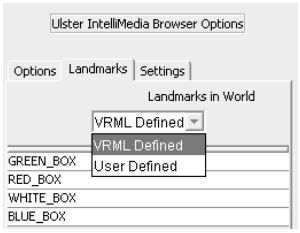


**Figure 5:** *Accessing landmarks via the interface.*

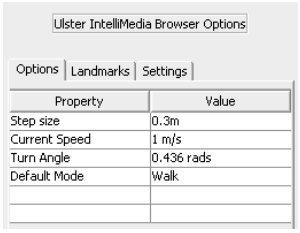


**Figure 6:** *Editable fields on the interface.*

the 2D interface (as a drop-down list, see Figure 5) where a double mouse click on the object's name will take the user to that object. There are, therefore, a number of ways in which the user can navigate through the VE using landmarks.

Earlier work [9] showed that a systematic concern of users with browsers was a lack of control over speed of movement. We have provided a number of ways to allow users to alter their speed. In walk mode, for example, a default *step length* is set to 1 m. To change the speed of movement, the user can either issue one of a number of spoken commands (e.g. "increase speed", "decrease speed") or can use the 2D interface to control a slider bar. The slider bar's values, which are also accessible by voice ("set speed to 5"), range from 0 to 10, and these values are then multiplied by the step length to increase the step length (walking faster) or decrease the step length (walking more slowly). We have chosen to implement speed controls in this way as the notion of a step length is universally known and the effect of lengthening or shortening the step length is predictable. Alternatively, users are able to set the speed (in metres/second) via an editable field on the interface (see Figure 6). This may be more useful where a VE representing a large area is used (e.g. in a virtual city).

Our earlier work also showed that users were concerned about the ability to *undo* navigation commands. This facility was seen as important not only when learning the system, when frequent mistakes were made, but also as a means of saving waypoints in order that steps could be retraced. We have adopted two conventions for saving waypoints that are somewhat analagous to how some word processors handle document version control. Similar to system-initiated *fast saves*, our system automatically saves all spoken navigation commands which can then be un-done in a linear fashion via a spoken "undo" command. Users can also initiate the storage of waypoints via a spoken "save waypoint" command at any time during navigation. These user-defined waypoints can then also be navigated in reverse. In a similar way to how the system handles landmark names, a drop-down box in the 2D interface allows user defined waypoints to be viewed; double-clicking on a particular waypoint takes the user to that position in the VE. In order that waypoints can be identified each waypoint is numbered. We are currently looking into ways to make this feature more effective by attaching landmark names to waypoints. It is likely, for example, that waypoints will be near to specific landmarks. This would enable users to store waypoints as they navigate from one landmark to another and then access these waypoints by referencing the landmark.

Finally, while our system has been built mainly to study the effectiveness of spoken dialogue when navigating through a VE, we have provided one inspection-related command which exploits the use of landmark names. For example, users can rotate an object (landmark) in the VE through a "rotate name" command. This causes the landmark in question to spin about its centre point through 360 degrees around the vertical $y$-axis, allowing the user a simple inspection of the object in-situ without altering his or her position in the VE. This may well prove to be more effective than using the mouse to move the user around the object, particularly when inspection of individual objects is required, such as in molecular visualization applications.

## 5. Implementation and Testing

Developing a spoken dialogue system is, thankfully, no longer a monumental task in itself, due to relatively recent advances in commercially available libraries. Our system is built on the Java Speech API and the implementation used here is IBM's ViaVoice system. Although this is a general-purpose speech recognition and speech synthesis system, it enabled a prototype of our system to be developed quite rapidly. For consistency and ease of development we use the Java 3D API combined with NCSA's Java3D Portfolio to handle different scene description formats. The system uses the standard Java runtime platform and Java Swing components provide the 2D interface. As described above, in order to support continuous actions the system is multi-threaded.

Users must first spend about 20 min training the system with the underlying IBM ViaVoice technology. Users are then asked to complete a basic training exercise using the VE navigation system, which does little more than introduce users to the range of commands recognized. A

small group of users (three), who had no previous experience of 3D browser software were given 17 different navigational and inspection tasks which were to be performed using (a) mouse only interaction and (b) speech and mouse interaction. Results show that the speech recognizer worked well with approximately 70% of commands recognized with the minimum recommended voice learning activity. This increased to nearly 90% when the users were encouraged to slow their speech and talk more clearly. The time taken to complete each task was recorded and users were also asked to rank the accuracy of each interaction style for each task. Results show that the speech and mouse interaction styles combined proved more accurate than only using the mouse for the majority of tasks. This was, we believe, due to the fact that precise navigational commands were possible with speech (e.g. "walk to the red car") but not possible using the mouse, as users had to position and orient themselves using mouse movements. However, using the mouse only proved to be marginally faster to complete the majority of tasks. This was confirmed by analysing users' comments, the most commonly reported negative comment being frustration with the length of time the speech system took to recognize a command and provide the updated visual. When walking, for example, the system takes about 2 seconds to implement the *stop* command. This delay may be due in part to the fact that the system currently uses the built-in ViaVoice large, general-purpose dictionaries. We are looking into the possibility of overcoming this problem by building a system with a smaller, bespoke dictionary.

## 6. Conclusions and Future Work

Efficient navigation remains challenging for many users and can be task-dependent. While mouse and keyboard interaction may be optimal for detailed interaction with virtual objects, they can be clumsy for navigational purposes. We set out to investigate the effectiveness of using spoken dialogue technology to support navigation in non-immersive VEs on desktop systems. While work remains to be done, a multi-modal system supporting speech and mouse interaction has been implemented using relatively standard technologies. Initial user response is positive, and results show that spoken dialogue systems for navigation can offer real advantages over traditional mouse and keyboard interaction. By speaking higher-level navigational commands, users are freed up from the cognitive and motor activities required to control hardware devices and therefore can devote more energy into achieving their goals. Furthermore, spoken commands (such as "walk to the red house") have the ability to be more goal-directed than a series of mouse movements.

The system is being refined to support more complex relations (e.g. topological and path relations) and also to support higher-level navigation and interaction with the VE and individual objects. For example, we are investigating the possibility of implementing Bolt's suggestions for a multi-modal graphics system from 1980 ("Put that there" [17]). We see potential for the integration of spoken dialogue systems with 3D graphics not just in navigation, but also in inspection and even scene building.

## 7. Acknowledgements

The authors would like to thank Siobhán McKinney for her work in testing the system. Thanks also to John Loughrey for software development on the project.

## References

1. J. Susanne and G.W. Furnas. Navigation in electronic worlds. In *CHI'97 Workshop, SIGCHI Bulletin*, 29(4), pages 44–49. 1997.
2. K. Kulwinder. Designing virtual environments for usability, PhD Thesis, Centre for Human-Computer Interface Design, City University, London, 1998.
3. S. Rushton and J. Wann. Problems in perception and action in virtual worlds. *Virtual Reality International'93*, 10:43–55, 1993.
4. K Stanney. Towards human-centered systems. *IEEE Computer Graphics and Applications*, 21–28, July/August, 1997.
5. M Mohageg, R Myers, C Marrin, J Kent, D Mott and P Isaacs. A user interface for accessing 3D content on the world wide web. In *Proceedings of CHI'96*, ACM Press, pages 466–472. 1996.
6. L. D. Miller. Metrics for Usability Standards in Computing (MUSiC): A usability evaluation of the Rolls-Royce virtual reality for aero engine maintenance system, Masters Thesis, University College, 1994.
7. T. Johnsgard. Fitt's law with a virtual reality glove and a mouse: effects of gain. In *Proceedings of Graphics Interface '94*, pages 8–15. 1994.
8. D.S. Tan, G.G. Robertson and M. Czerwinski. Exploring 3D navigation: combining speed-coupled flying with orbiting. In *Proceedings ACM CHI 2001*, pages 418–425. 2001.
9. H.M. Sayers, S. Wilson, W. Myles and M.D.J. McNeill. Usable interfaces for virtual environment applications on non-immersive systems. In *Proceedings Eurographics UK Conference*. 2000.
10. N.G. Vinson. Design guidelines for landmarks to support navigation in virtual environments. In *Proceedings of ACM CHI'99*, pages 142–149. 1999.

11. A. van Ballegooij and A. Eliens. Navigation by query in virtual worlds. In *ACM Web3D Conference 2001*, Paderbon, Germany, pp. 77–83. 2001.

12. V. Tanriverdi and R.J.K. Jacob. Interacting with eye movements in virtual environments. In *Proceedings ACM CHI 2000*, pages 265–272. 2000.

13. P. Mc Kevitt (ed.). *Integration of Natural Language and Vision Processing (Volumes I-IV)*. Kluwer Academic Publishers, 1995/96.

14. D. Golightly, K.S. Hone and F.E. Ritter. Speech interaction can support problem solving. In M. Angela Sasse and Chris Johnson (eds), *Human-Computer Interaction, INTERACT'99*, IOS Press, pages 149–155. 1999.

15. M. Billinghurst, J. Savage, P. Oppenheimer and C. Edmond. The expert surgical assistant: an intelligent virtual environment with multimodal input. In S. Weghorst, H.B. Sieberg and K.S. Morgan (eds), *Proceedings of Medicine Meets Virtual Reality IV*, pages 590–607. 1995.

16. D. Weiner and S.K. Ganapathy. A synthetic visual environment with hand gesturing and voice input. In *Conference on Human Factors in Computing Systems (CHI'89)*, pages 235–240. 1989.

17. R.A. Bolt. Put-That-There: voice and gesture at the graphics interface. In *ACM SIGGRAPH Proceedings 1980*, pages 262–270. 1980.

18. P. Mc Kevitt. Visions for language. In *Proceedings of the Workshop on Integration of Natural Language and Vision processing*, Twelfth American National Conference on Artificial Intelligence AAAI-94, pages 44–57. 1994.

19. P. Mc Kevitt. SuperinformationhighwayS "Sprog og multimedier" (speech and multimedia). In Tom Bröndsted and Inger Lytje (eds), Aalborg University Press, pages 166–183. 1997.

20. D. Partridge. *A new guide to Artificial Intelligence*. Ablex Publishing Corporation, 1991.

21. E. Rich and K. Knight. *Artificial Intelligence*. McGraw-Hill, 1991.

22. D. Dennett. *Consciousness explained*. Penguin, 1991.

23. S.M. Kosslyn and J.R. Pomerantz. Imagery, propositions and the form of internal representations. *Cognitive Psychology*, (9):52–76, 1977.

24. Z. Pylyshyn. What the mind's eye tells the mind's brain: a critique of mental imagery. *Psychological Bulletin*, (80):1–24, 1973.

25. W. Wahlster. One word says more than a thousand pictures: On the automatic verbalization of the results of image sequence analysis. *Bereich Nr. 25*. Universität des Saarlandes, 1988.

26. D. Waltz. Understanding line drawings of scenes with shadows. *The psychology of computer vision*, P.H. Winston (ed), McGraw-Hill, pages 19–91. 1975.

27. G. Herzog and G. Retz-Schmidt. Das system SOCCER: simultan Interpretation und naturalichsprachliche. *Beschreibung zeitveranderlicher Szenen Sport und Informatik*. J. Perl (ed), Hofmann, 1990.

28. B. Neumann and H.J. Novak. NAOS: ein system zur naturalichsprachlichen. *Beschreibung zeitveranderlicher Szenen Informatik, Forschung und Entwicklung*, 1(1):83–92, 1986.

29. G. Retz-Schmidt. Recognizing intentions, ineractions, and causes of plan failures. *User Modelling and User-Adapted Interaction*, (1):173–202, 1991.

30. G. Retz-Schmidt and M. Tetzlaff. Methods for the intentional description of image sequences. *Bereich Nr. 80*. Universität des Saarlandes, Germany, 1991.

31. G. Herzog, C.K. Sung, E. André, W. Enkelmann, H.H. Nagel, T. Rist and W. Wahlster. Incremental natural language description of dynamic imagery. In C. Freksa and W. Brauer (eds), *Wissenbasierte Systeme. 3*, Springer-Verlag, pp. 153–162. 1989.

32. E. André, G. Herzog and T. Rist. On the simultaneous interpretation of real-world image sequences and their natural language description: the system SOCCER. In *Proceedings of the 8th European Conference on Artificial Intelligence*, pages 449–454. 1988.

33. W. MaaS, P. Wiziniski and G. Herzog. VITRA GUIDE: Multimodal route descriptions for computer assisted vehicle navigation. *Bereich Nr. 93*. Universität des Saarlandes, 1993.

34. M. Maybury (ed.). *Intelligent multimedia interfaces*. AAAI Press, Menlo Park, CA, 1993.

35. M. Maybury and W. Wahlster (eds.). *Readings in intelligent user interfaces*. Morgan Kaufmann Publishers, 1998.

36. G. Carenini, F. Pianesi, M. Ponzi and O. Stock. Natural language generation and hypertext access. In *IRST Technical Report 9201-06*. Instituto Per La Scientifica E Tecnologica, Loc. Pant e Di Povo, I-138100 Trento, Italy, 1992.

37. O. Stock. Natural language and exploration of an information space: the ALFresco Interactive system. In

*Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI-91)*, pages 972–978. 1991.

38. E. André and T. Rist. The design of illustrated documents as a planning task intelligent multimedia interfaces. In M Maybury (ed), AAAI Press, Menlo Park, CA, pages 75–93. 1993.

39. W. Wahlster, E. André, W. Finkler, M.J. Profitlich and T. Rist. Plan-based integration of natural language and graphics generation. *Artificial Intelligence, Special issue on natural language generation*, (63):387–427, 1993.

40. M. Maybury. Planning multimedia explanations using communicative acts. In *Proceedings of the Ninth American National Conference on Artificial Intelligence (AAAI-91)*, pages 14–19. 1991.

41. T. Broendsted, P. Dalsgaard, L.B. Larsen, M. Manthey, P. Mc Kevitt, T.B. Moeslund and K.G. Olesen. A platform for developing intelligent multimedia applications. In *Technical Report R-98-1004*. Center for Person Kommunikation (CPK), Institute for Electronic Systems (IES), Aalborg University, Denmark, 1998.

42. T. Broendsted, L.B. Larsen, M. Manthey, P. Mc Kevitt, T.B. Moeslund and K.G. Olesen. The IntelliMedia WorkBench: a generic environment for multimodal systems. In Robert H. Mannell and Jordi Robert-Ribes (eds), *Proceedings of the Fifth International Conference on Spoken Language Processing (ICSLP-98)*, (2), pages 273–276. 1998.

43. T. Broendsted, L.B. Larsen, M. Manthey, P. Mc Kevitt, T.B. Moeslund and K.G. Olesen. CHAMELEON: a general platform for performing intellimedia. In P. Mc Kevitt, S. O'Nuallain and C. Mulvihill (eds), *Language, Vision and Music*, AiCR, John Benjamins Publishing, Amsterdam, pages 79–96. 2002.

44. T. Broendsted, P. Dalsgaard, L.B. Larsen, M. Manthey, P. Mc Kevitt, T.B. Moeslund and K.G. Olesen. The IntelliMedia WorkBench - an environment for building multimodal systems. In *Advances in Cooperative Multimodal Communication: Second International Conference, CMC'98*, pages 217–233. 1998.

45. T. Broendsted, P. Dalsgaard, L.B. Larsen, M. Manthey, P. Mc Kevitt, T.B. Moeslund and K.G. Olesen. Developing intelligent multimedia applications. *Multimodality in Language and Speech Systems (MiLaSS)*, B. Granstrom, I. Karlsson and D. House (eds), Kluwer Academic Publishers, pages 149–171. 2002.

46. P. Mc Kevitt and P. Dalsgaard. A frame semantics for an IntelliMedia TourGuide. In *Proceedings of the Eighth Ireland Conference on Artificial Intelligence (AI-97), volume 1*, University of Ulster, Northern Ireland, pages 104–111. 1997.

47. G. Rickheit and I. Wachsmuth. Collaborative Research Centre "Situated Artificial Communicators" at the University of Bielefeld. *Integration of Natural Language and Vision Processing, volume IV, Recent Advances*, P. Mc Kevitt (ed), Kluwer Academic Publishers, pages 11–16. 1996.

48. K.R. Thörisson. Communicative humanoids: a computational model of psychosocial dialogue skills, Ph.D. thesis, Massachusetts Institute of Technology, 1996.

49. K.R. Thörisson. Layered action control in communicative humanoids. In *Proceedings of Computer Graphics Europe '97*. 1997.

50. Naoyuki Okada. Integrating vision, motion and language through mind. *Integration of Natural Language and Vision Processing, volume IV, Recent Advances*, P. Mc Kevitt (ed), Kluwer Academic Publishers, pages 55–80. 1996.

51. N. Okada. Integrating vision, motion and language through mind. In *Proceedings of the Eighth Ireland Conference on Artificial Intelligence (AI-97), volume 1*, University of Uster, Northern Ireland, pages 7–16. 1997.

52. A. Waibel, M.T. Vo, P. Duchnowski and S. Manke. Multimodal interfaces. In P. Mc Kevitt (ed), *Integration of Natural Language and Vision Processing, volume IV, Recent Advances*, Kluwer Academic Publishers, pp. 145–165. 1996.

53. M. Denis and M. Carfantan (eds.). Images et langages: multimodalit/e et modelisation cognitive. *Actes du Colloque Interdisciplinaire du Comit/e Nationa de la Recherche Scientifique*. Salle des Conférences, Siége du CNRS, Paris, 1993.

54. A. Pentland (ed.). Looking at people: recognition and interpretation of human action. In *IJCAI-93 Workshop (W28) at The 13th International Conference on Artificial Intelligence (IJCAI-93)*. Chambéry, France, 1993.

55. D. Bowman, D. Koller and L. Hodges. Travel in immersive virtual environments: an evaluation of viewpoint motion control techniques. In *Proceedings of the 1997 Virtual Reality Annual International Symposium (VRAIS)*, pages 45–52. 1997.

56. R.P. Darken and J.L. Silbert. A Toolset for Navigation in Virtual Environments. In *Proceedings of UIST '93*, pages 157–165. 1993.

57. D.D. Salvucci and J.R. Anderson. Intelligent Gaze-added interfaces. In *Proceedings ACM CHI2000*, pages 273–280. 2000.