1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
7.1
8.0
9.0
10.0

1 of 1

SAN94-0670C
Conf-9406136--1

USING VOICE INPUT AND AUDIO FEEDBACK TO ENHANCE THE REALITY OF A VIRTUAL EXPERIENCE[1]

Nadine E. Miner
Member of the Technical Staff
Sandia National Laboratories
Albuquerque, NM

Abstract

Virtual Reality (VR) is a rapidly emerging technology which allows participants to experience a virtual environment through stimulation of the participant's senses. Intuitive and natural interactions with the virtual world help to create a realistic experience. Typically, a participant is immersed in a virtual environment through the use of a 3-D viewer. Realistic, computer-generated environment models and accurate tracking of a participant's view are important factors for adding realism to a virtual experience. Stimulating a participant's sense of sound and providing a natural form of communication for interacting with the virtual world are equally important. This paper discusses the advantages and importance of incorporating voice recognition and audio feedback capabilities into a virtual world experience. Various approaches and levels of complexity are discussed. Examples of the use of voice and sound are presented through the description of a research application developed in the VR laboratory at Sandia National Laboratories.

1 Introduction

Voice command input and audio feedback are two important ways in which the realism of a virtual experience can be enhanced. These VR interaction methods are also a practical alternative to the use of methods such as virtual menu systems, physical buttons and hand gesture recognition. Interaction through the use of verbal commands and audio feedback is a natural and accepted form of human communication and therefore provides a more intuitive human-computer interface paradigm than these other methods.

Voice recognition can be incorporated into a VR simulation to provide the participant with the ability to interact with the virtual environment through the use of verbal commands. A voice recognition interface to a virtual environment can provide a natural, easy-to-use interface for all users of the system, including those with disabilities. Voice interfaces to virtual worlds have not been extensively researched for application within the VR community. There are many advantages to using a voice interface in a VR world which are outlined in Section 2. Also discussed in this section are some issues to be considered when integrating a voice interface into a VR system.

Audio feedback is an important element of the human communication loop. Many VR researchers have experimented with sound to communicate perceptual information to a VR participant, among them are: [Cohen 92], [Scarborough 92] and [Wenzel 90]. Section 3 discusses the many ways audio feedback can be used to add realism to a virtual world. A process for incorporating audio feedback and sound sequence generation are is discussed in section 3.

Voice input and sound feedback have been successfully incorporated into several VR applications. Yonekura has developed a voice activated system with sound feedback for moving objects in a 3-D virtual world [Yonekura 93]. In the VR systems developed at Sandia's VR laboratory, we have found that a wide range of users become adept in the use of the systems very quickly, largely due to the intuitive nature of the interactions. Section 4 contains a description of one of these Sandia-developed applications in which the user controls a complex robotic system.

[1] This work was performed at Sandia National Laboratories supported by the US Dept. of Energy under contract number DE-AC04-94ALS85000.

Presented at the IMAGE VII Conference, Tucson, AZ 12-17 June 1994.

Section 5 discusses some of the many enhancements which will be required of voice and sound systems to further increase the realism of the participant's interactions with the virtual environment.

2 Voice Interface Advantages and Development

There are several advantages to using voice command input in a VR system. Users can be quickly trained in the use of the VR system due to the intuitive nature of the voice interaction and the ability to custom-design application languages. The usually small size of an application specific vocabulary further simplifies the task of learning the interface. Often times, voice commands are more natural to use than hand gestures or virtual menu selections, and are therefore easier for participants to learn. The use of voice allows the participant to focus on the VR experience itself instead of a more cumbersome interaction mechanism. Assuming that the recognition rate is sufficiently high, voice command input tends to be less tiring for participants to use. For example, it's much faster and easier to give the voice command "increase speed by 10" then to give a hand gesture, or to go through a series of virtual pull-down menus to achieve the same result. Voice command interaction is especially useful for users with disabilities where hand gesture input may not be an option.

In Sandia's VR laboratory, we have implemented voice command interfaces for several different VR applications. We have chosen to limit the command vocabulary, or language, to application-specific words or phrases. The user inputs the verbal commands, which are recognized by the system, to achieve a particular result. A limited vocabulary also has the advantage of a higher recognition rate because the language set is significantly reduced. We are also able to add a hierarchical structure to the vocabulary which further increases the recognition rate and reduces the number of "misfires", or mis-recognized commands.

There are many off-the-shelf, speaker independent voice recognition systems available which provide the capabilities of designing and training custom vocabularies. We are currently using a PC-based system called DragonWriter™ by Dragon Systems. Speaker independent performance can be obtained by training the vocabulary on a sample set of voices and "building" a voice template for a custom-designed vocabulary. The success rate of the voice recognition is dependent on the vocabulary phrases and the set of training voices selected, as is discussed in the next section. DragonWriter™ outputs user-defined ASCII characters when voice commands are recognized. We have developed an interface between the DragonWriter™ system and our VR applications which run on Silicon Graphics, Inc. (SGI) platforms. This software provides the link between the voice recognition system and the resulting action within the VR environment. We have found that it is very important to combine audio feedback with the voice command interface to provide the user with both visual and auditory confirmation that voice commands have been received and recognized; this will be discussed in more detail in Section 3.

Voice Interface Issues to Consider

Voice recognition systems available today perform best when used repeatedly with a single user, in a controlled environment, with non-continuous speech, and with a limited vocabulary. Markowitz discusses the many issues associated with continuous, speaker-independent Automatic Speech Recognition (ASR) systems and the progress that needs to be made in this area [Markowitz 93]. When using discrete commands and a limited vocabulary, reasonable success can be achieved, even with users who have not trained on the voice system and when the environment is noisy. However, with today's systems, the success of the voice recognition depends largely on the choice of vocabulary words and phrases themselves. There is a trade-off to be made between the length of command phrases and the success rate of the recognition system. We have found that very short commands result in a large misfire rate. On the other hand, long command phrases tend to have a low recognition rate

because few people say long phrases similarly; even a single speaker has difficulty saying a long phrase the same way twice. Also, some consonant and vowel sounds have more variation among different speakers. For example, in one VR application, the word "gripper" is used. This word has a low recognition rate in comparison to the other words in the vocabulary, but also a very low misfire rate. This means that, although the word is distinct, there is too much variation in the pronunciation of the word among speakers. We have found that two or three word phrases have the highest recognition rate and lowest misfire rate. The design of a custom vocabulary takes experience and/or study of human sound pronunciations.

The recognition rate is also dependent on the particular voices involved in the training set. We typically use a training set of 10 voices, 5 male and 5 female. We do not use heavily accented voices nor voices with extremes in volume or pitch. Again there are trade-offs between the number of voices used in the training set and the recognition rate achieved. If too many voices are used to train the system, the recognition rate is degraded. If too few voices are used in the training set, the recognition rate will also be low, except for those upon whose voices the system was trained, because there will not be enough variation in the voice template . We have found that systems trained with a 5/5 male/female mix has a low success rate with recognizing very deep pitch voices. This might be alleviated by adding voices to the training set which are more varied in pitch, but this would risk lowering the recognition rate for the average speaker. In order for voice recognition systems to achieve consistently greater performance, the context specific meaning of spoken words will have to be able to be determined. This is an active area of research known as natural language processing. Details of natural language processing are beyond the scope of this paper.

3 Use of Audio Feedback

Audio feedback can provide several useful benefits in a VR system. Speech synthesis can provide a means of acknowledging the recognition and execution of voice input commands. This follows well with the natural and expected method of human communication. Additionally, the participant can be guided through the use of a virtual system by synthesized audio instructions. Real-time help, system status and novice user instructions can be provided through the audio system as the participant moves through the VR world. Additionally, if headphones are used, the ambient noise from the real world is reduced, thus serving to further increase the participant's sense of immersion. We have found these audio feedback techniques to be very useful in rapidly training participants in the use of our VR systems; users become comfortable with the system very quickly and have a strong sense of immersion in the virtual environment.

Sound effects, including 3-D sound, can be used to greatly enhance the realism of the VR environment. For example, if a VR simulation consisted of a jungle environment, the inclusion of jungle noises and sound effects would make the scenario much more realistic. Lehnert and Blouert discuss the many aspects involved in a "virtual sound environment" [Lehnert 91]. The incorporation of 3-D sound will serve to further increase the realism of the audio feedback in a virtual experience. With 3-D sound, simulation of passing cars or trains can be realistically created. Begault discusses some of the difficulties associated with 3-D sound generation and some of the preliminary results obtained at the Auditory Lab at NASA Ames Research Center [Begault 91].

Audio effects can also be used in place of haptic feedback to provide the participant with an indication of valid or invalid actions. Because of the difficulties involved with simulating realistic haptic, or force feedback [Shimoga 93], sound can be used as a straight forward guide to the user in performing a task. NASA Ames Researchers demonstrated this use of audio feedback when they used sounds to guide a participant in placing an object in a specific location with a particular orientation.

Sound Interface Issues to Consider

Quality sound effects can be very tedious, time consuming, and expensive to create. However, sound effect libraries which contain literally thousands of sound effects are becoming more widely available. Probably the best known and most comprehensive collection of sound effects has been created by Sound Ideas. These sound effects have been used by video, film and music professionals worldwide [Mc Pherson 93]. Several companies have been licensed to sell subsets of the Sound Idea collection, so application specific effects can be purchased without requiring a user to purchase the entire library.

The applications developed in Sandia's VR lab use a combination of sound effects and preprogrammed voice sequences to responded to user commands or actions. Voice feedback sequences can be created in several ways. Human voice sequences can be digitized and stored in computer memory for later playback. This method has the advantage of providing realistic, natural voice feedback, but also requires a lot of computer memory capacity. Alternatively, voice feedback can be synthesized by products such as the Sound Blaster™ by Creative Labs, Inc. This approach has the disadvantage of being less realistic sounding than human digitized voices, but only requires storage space for the ASCII strings and commands to the voice synthesizer.

Long sequences of sounds required for background sound effects present a different type of problem. Storage of very long sequences in memory is typically cost prohibitive, especially considering that VR experiences can last many minutes. Thus, shorter background sound sequences must be replayed for the duration of a VR experience. If the background sounds are repetitive and are looped over, the resulting background sound will seem artificial. Thus, sound samples without distinct, repetitive noises are required to generate long sound sequences. For added realism, distinctive sounds should be mixed in at random intervals. Take for example the sound of a busy street: you would want the background sound of a steady hum of traffic to loop over. Occasional horns blowing, sirens blaring, or bells ringing could be mixed in randomly to obtain a realistic sound sequence.

Generation of realistic sound effects, as discussed above, these leads to the need for rather sophisticated software and/or hardware sound processing tools. We are currently using the embedded sound capabilities of the SGI Indigo™ system, since this is the platform on which our VR applications are based. The Indigo™ system has sound digitizing, storage, and playback capability, but does not currently have any advanced software sound processing tools, such as mixing or blending. Thus, we are investigating the use of other hardware and software sound systems on the market which offer more sound processing capabilities. Most of these systems use the MIDI (Musical Instrument Digital Interface) standard for generating and controlling sounds. By using the MIDI standard, VR applications can control hundreds of different devices which produce music or manipulate previously sampled sounds. Two examples of commercially available sound synthesis products based on this technology are the Sound Cube™ from Visual Synthesis, Inc. and the Paradigm™ sound system from Paradigm Simulations. Sound localization for 3-D sound requires additional hardware. Two well known sound localization products are the Convolvotron by Crystal River Engineering, and the Focal Point 3D Audio system. These products work in conjunction with sound synthesis systems to simulate 3-D sound.

4 Using Voice Interface and Audio Feedback for Robot Control

All of the voice driven, sound feedback VR systems developed in Sandia's VR laboratory utilize application-specific, custom designed vocabularies along with pre-digitized voice feedback sequences and sound effects. These applications are highly-structured in terms of the vocabularies which they are capable of recognizing. One such VR system allows a participant to interact with a complex robotic system from the VR interface and will eventually allow participant control of the actual robot located in a remote location [Miner 93]. The robot system consists of two six-degree-of-

freedom robots, one mounted on the other, along with various tools and sensors [Davies 93]. Fig. 1 shows a block diagram of the system interface configuration for this VR system. Fig. 2 shows the VR model of the robot workcell.

Initially, it was intended that hand gestures and arm movements would be interpreted to directly control the robot motion. However, preliminary experiments showed that there was no intuitive mapping between human hand/arm motions or gestures and this complex robotic system. Other goals of this application were to permit non-robotic experts to operate the system with minimum training and for extended periods of time without becoming fatigued. For these reasons, we implemented a voice controlled interface with appropriate real-time audio feedback to accomplish the application goals. The voice commands could be easily learned and would initiate sequences of robot instructions, the details of which the novice user need not be concerned. Audio feedback was integrated into the system to provide the user with confirmation of commands received. Additionally, audio feedback was used to provide guidance to the user during training and use of the system in the form of help and status facilities.

A speaker independent, hierarchical vocabulary was designed for this application. As words and phrases are recognized, ASCII characters are output by the PC to the SGI workstation through a serial link. Voice commands are processed by the SGI once per second so that voice command response is timely without too severely impacting the graphics update rate. Once the commands are confirmed, appropriate sets of robot instructions are generated and simulated in the VR environment for the user to observe.

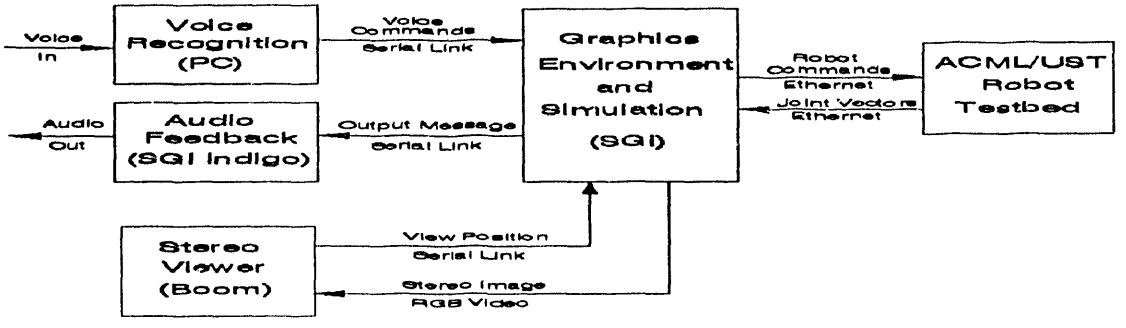

Figure 1: VR System Block Diagram

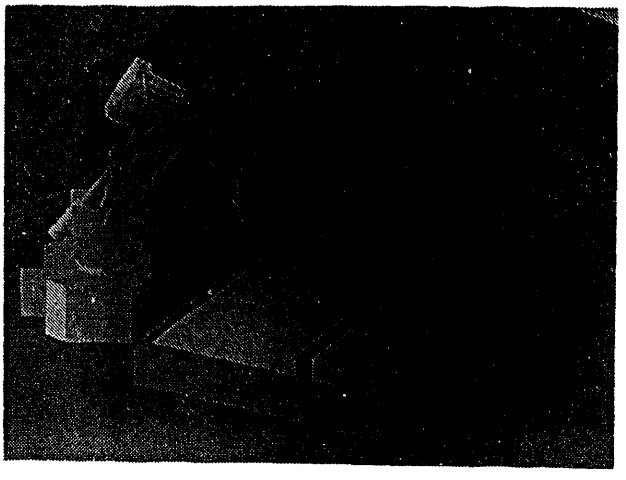

Figure 2: VR model of UST Robot workcell

The audio feedback capability is provided by an Indigo 2 workstation via a socket server. Message pointers are sent to the audio server to initiate appropriate audio feedback messages depending on the users actions and commands. Table 1 gives an example of the voice commands and the corresponding audio feedback sequences used in this application. As an example, the operator will speak the command "Get Cutter". This command is recognized by the voice system and, if valid at this stage, is translated into a set of commands which drive the robot simulation. Audio feedback informs the user whether the command is valid or not. We have observed an ~85% recognition rate using the DragonWriter™ system on this vocabulary set, including users with noticeable foreign accents, using a language template developed with a training set of 5 male and 5 female speakers.

Voice Command	Audio Feedback Valid Command	Audio Feedback Invalid Command	Simulated Action valid commands
get cutter	*getting cutter tool*	*cannot get cutter*	cutter tool grasped
cut pipe	*executing cut operation*	*cannot cut pipe*	cut pipe invoked
open gripper	*opening gripper*	*cannot open gripper*	gripper opened

Table 1: Examples of Voice Commands and Audio Responses

5 Conclusions and Future Work

Overall, we have found the use of a voice interface and audio feedback to be very beneficial in a variety of VR applications. Participants become comfortable with the use of these systems very quickly and extended use is not strenuous. Using structured vocabularies and pre-digitized audio feedback sequences, the participant is able to interact with the VR system in a more natural way than other methods might provide. However, the user interactions with the VR applications discussed are limited to pre-determined voice commands and responses, so a completely natural, conversational interaction between the user and the VR system is not possible at this time. A significant amount of system intelligence would have to be incorporated into the system to recognize conversational speech and to be able to respond dynamically with appropriate audio feedback. As voice recognition, audio synthesis, and VR technologies progress, it will become desirable to provide this type of natural language interaction where the VR system can respond dynamically to context-dependent voice inputs. Niimi describes some of the issues and problems that must be solved to realize this type of "friendly dialogue" system [Niimi 93]. Among the problems to be solved are: improved word recognition rate over continuous, speaker-independent speech; context dependent recognition of words, dynamic voice synthesis so that realistic sounding speech can be generated on-the-fly instead of being limited to canned or preprogrammed responses; and the ability to "exchange initiative" between the user and the machine which would allow the computer to offer information with which the user may not have specifically requested. Even with the currently limited capabilities of existing systems, however, voice input and audio feedback should be considered to be vital components for a realistic VR experience.

References

[Begault 91] Begault, D. R., "Challenges to the Successful Implementation of 3-D Sound", Journal of Audio Engineering Society, Vol. 39, No. 11, Nov. 1991.

[Cohen 92] Cohen, M., "Integrating Graphic and Audio Windows", *Presence, Vol. 1, No. 4,* pp. 468-467, Fall 1992.

Crystal River Engineering, *The Convolvotron: Synthetic 3D Audio.* Sales Brochure. 12350 Wards Ferry Road, Groveland, CA. 1990.

Dragon Systems, Inc., *The DragonWriter Development System.* Brochure. 320 Nevada Street, Newton, MA 02160.

[Davies 93] Davies, B. R., "Remediating Hazardous Waste Robotically Using a High-Level Control System and Real-Time Sensors", Proc of the SPIE International Symposium on Optical Tools for Manufacturing and Advanced Automation, Sept. 1993.

Focal Point 3-D Audio, *3-D Audio for the Macintosh II.* Sales Brochure. 1402 Pine Avenue, Suite 127, Niagara Falls, NY 14301. 1991.

[Lehnert 91] Lehnert H., Blouert J., "Virtual Auditory Environment", Proceedings of 1991 ICAR, Pisa, Italy, 1991.

[Niimi 93] Niimi, Yasuhisa, "How Might One Comfortably Converse with a Machine?", IEICE Transactions on Information & Systems, Vol. E76-D, No 1, pp. 9-16, Jan 93.

[Markowitz 93] Markowitz, Judith, "The Power of Speech", AI Expert, Vol. 8, No. 1, pp. 28-32, Jan 1993.

[Mc Pherson 93] Mc Pherson, M., "Speed-of-Sound", Aware, Inc., Product Literature, 1993.

[Miner 93] Miner, N., Stansfield, S., "An Interactive Virtual Reality Simulation System for Robot Control and Operator Training", submitted to ICRA 1994.

Paradigm Simulation, Inc., *Audio Works Technical Overview.* Sales Brochure. 15280 Addison Road, Suit 120, Dallas, TX 75248. 1993.

[Scarborough 92] Scarborough, E., Brandt, J., et. al, "A Prototype Visual and Audio Display", *Presence, Vol. 1, No. 4,* pp. 459-467, Fall 1992.

[Shimoga 93] Shimoga, K., "A Survey of Perceptual Feedback Issues in Dexterous Telemanipulation: Part I. Finger Force Feedback", Proc of IEEE Virtual Reality Annual Intl Symposium (VRAIS), pp. 263-270, Sept. 1993.

Visual Synthesis, Inc., *Sound Cube,* Sales Brochure. 4126 Addison Road, Fairfax, VA 22030, 1993.

[Wenzel 90] Wenzel, E. M., Stone, P. K., et. al, "A System for Three-Dimensional Acoustic "Visualization" in a Virtual Environment Workstation", Proceedings of the First IEEE Conference on Visualization, pp. 329-337, 1990.

[Yonekura 93] Yonekura, T., Nobuyuki, A., and Watanabe, Y., "ASPECT: Audio SPatial Environment for CommunicaTion - as a Three Dimensional Auditory Interaction Tool", Proceeding of the IEEE Virtual Reality Annual International Symposium (VRAIS), pp. 263-270, Sept. 1993.

DATE
FILMED
5/9/94
END