Communication with VR Training System using Voice and Behavior

Kazuaki Tanaka, Tomoaki Ozaki and Norihiro Abe
Faculty of Computer Science and System Engineering,
Kyusyu Institute of Technology
Iizuka-shi, 820-8502 Japan
kazuaki@mse.kyutech.ac.jp
Hirokazu Taki
Faculty of System Engineering, Wakayama University, Japan

Abstract

In this research the training system using a virtual reality system was developed to instruct assembly/disassembly of mechanical parts to a user. A bidirectional interface system is realized that permits a user and the system to communicate each other using verbal and nonverbal information. When a user has questions in the process of operation, he can ask or give an order to the system that is an instructor using a spoken language and nonverbal behavior such as pointing action. A model of the instructor, an avatar is rendered in the virtual environment, he replies to questions or commands from a user. While an avatar uses a spoken language and can show instruction and operation of virtual parts with his behavior. Not only the synchronized recognition of voice and behavior of a user, but also the synchronization mechanism of the speech synthesis and the behavior generation of an avatar were stated clearly.

Keywords: Verbal/Non-verbal Communication, Training System, Assembly of Mechanical Parts

1. Introduction

We have reported several papers on the training system in mechanical assembly/ disassembly domain using a virtual reality system [1~3]. This training system is different from the traditional one using a mouse and a keyboard. It can watch the behavior of a user and instruct a right way when his action is wrong. But this system is not able to know the intention of the user definitely because no voice interaction facility is provided with the user. In other words, only by watching the human nonverbal behavior, the system can't completely detect the human intention or hesitation [2].

In communication between human beings, a spoken language becomes important besides the nonverbal behavior. So, in this research we propose a training system with verbal/nonverbal communication facility between human being and a computer system. In an assembly/disassembly training system, a user is permitted to get into a virtual environment in which a virtual machine is rendered and to perform a simulation of assembly/disassembly operation. When a user has questions in the process of operation, he can ask or give an order to the system which is an instructor in a spoken language. In this system, a model of the instructor, an avatar is rendered in the virtual environment, he replies to questions or commands from a user. While an avatar uses a spoken language and can show instruction and operation of virtual parts with his behavior.

In this system, both spoken language and nonverbal behavior can be input at the same time in order to realize verbal/ nonverbal communication. A virtual reality system can be brought closer to a real environment by using this interface. For example, we can ask a question or issue an order about an object to the system using a spoken language pointing at the object with a data glove. The system developed in this research permits an avatar to perform communication with a user using a spoken language information and a non-verbal information.

This time, as the field of the application of a bidirectional communication using verbal / non-verbal information, we selected the field of assembly / disassembly of mechanical parts. But we think this system applicable to various interface between human and machine

In this research, we proceed the research to attain a big aim to bring the communication between human and machine close to that between human beings.

2. System configuration

2.1. Hardware organization

The system consists of the computer which builds a virtual reality system, a microphone for a user to perform voice input, and 3-dimensional position sensors and data gloves for a user to input non-verbal behavior. General drawing is shown in Figure 1.

2.2. Continuous speech recognition parser (JULIAN)

This research used JULIAN which Prof. Doshita's research laboratory in Kyoto University developed as a speech recognition software. JULIAN is a recognition parser performing continuous speech recognition on the basis of a finite state grammar (DFA). It begins to look for the most plausible word list based on a given DFA for voice input from the microphone (continuous speech to make a pose with gap) and outputs it as a character string. DFA is made from vocabulary and the syntax rule that a user registered.

2.3. OpenInventor

To build a virtual reality system, three-dimensional surface models are used. A three-dimensional graphics library, OpenInventor [4] of SGI Company is used.

3. Assembly training system

3.1. System configuration

This system consists of 3 parts including avatar unit, spoken language processing unit, and non-verbal behavior analysis unit.

In a spoken language processing unit after voice input from a user is processed through a speech recognition and natural language processing sub-system, the result is given to an avatar unit.

In the nonverbal behavior analysis unit, hand position and attitude of the user are analyzed and the result is transmitted to the avatar unit. The avatar unit estimates the information sent and takes the factual knowledge of the virtual machine described in the system to make an appropriate response to a user.

We summarize the main facility of each part in Figure 2. We describe each function in detail later.

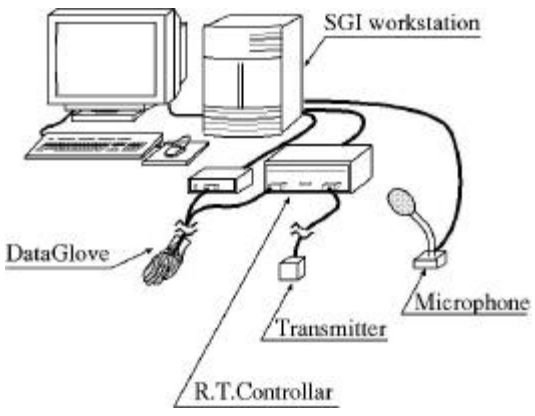


Figure1. Hardware organization

3.2. Verbal/ non-verbal interface

In this system, we used the interface that a user could input spoken language and non-verbal behavior simultaneously. Consequently, a user has only to utter toward a microphone in case issuing voice input without any keyboard action. The operation method peculiar to this interface is shown in the following.

i. With a traditional interface, the unique name must be used in order to distinguish the object from others. But when there are many same objects like mechanical parts, it is difficult to designate one of them using the name. This is, however, easily realized simply by pointing or grasping the object. A user can speak to the system by inputting the spoken language such as "Install this part on that." while pointing at the two objects with a data glove. A user will need not memorize the identifier of object parts by admitting the use of the directive. How to make correspondence between terms meaning instruction such as 'this', 'that' and the behavior like a pointing action will be fully described in 6.2.

ii. A user is able to order the avatar to do assembly operation. If we should want to interrupt the operation while the avatar is executing an assembly operation, we could have the avatar suspend the operation by issuing a phrase or sentence that means the suspension of the operation.

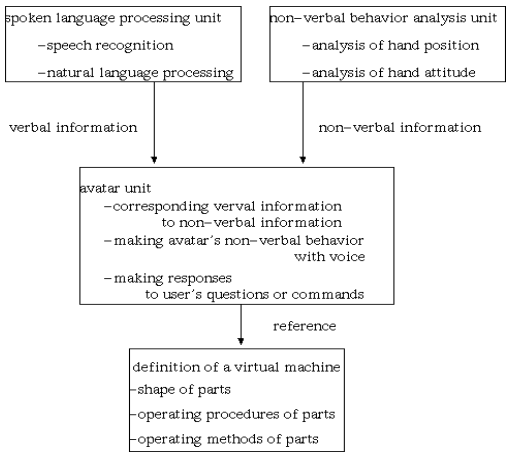


Figure 2. System configuration

3.3. Definitions of operating procedure

In this system, the operating procedure (AND/OR procedure) is defined with an AND/OR graph as shown in Figure 3. Hereafter, we call the part a mvobject which has a component to be moved after a user has selected it with a data glove or voice input, and call the partner part a basic part into which the mvobject is installed. Each node in the AND/OR graph shown in Figure 3, for examples START, END and points from 1 to 8, expresses an assembly status of the give assembly.

Assembly operation along an arc of the graph (operation) is necessary in order to change the state of the assembly. In operation, operating instruction and the object parts (mvobject, basic part) are described.

All nodes of the AND/OR procedure shown in Figure 3 consists of OR nodes. In other words, a user has only to sequentially follow the graph from the upper part toward the lower part. For example, assembly procedures such as [START - 1-5 - END], [START - 3-8 - END] are right procedures.

3.4. Definitions of mechanical part

In this system mechanical parts are defined with a Scene Graph [4] as shown in Figure 4. A **MyParts** is data node. A part name and a part number are described in the **MyParts**. A part name corresponds to the voice input from a user. A part number is used for describing the object part in the AND/OR procedure.

3.5. Assembly method

A basic assembly method of virtual part should be explained here. To make clear an assembly method and to help the operation of a virtual part, an arrow is attached to each portion of the part to be mated with another part as shown in the Figure 5. The direction coincides with that of the assembling operation. If the following conditions are met, then the operation is automatically finished at the final state; A user is moving each part to the direction of an arrow. The

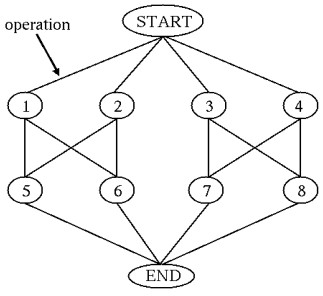
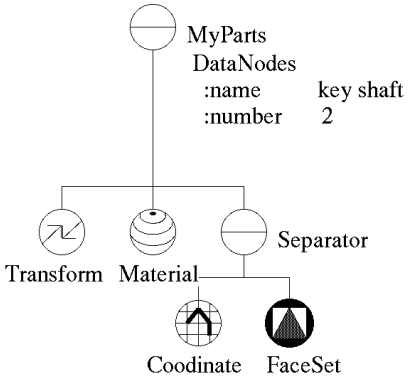


Figure 3. Operating procedure based on AND/OR graph



MyParts : definition of a part name and a part number
Transform : definition of a part position
Coordinate, FaceSet : definition of a part shape

Figure 4. Definition of mechanical part

roots of arrows attached to the parts get closer each other.

And in this system, the operating procedure (AND/OR procedure) is defined with an AND/OR graph [5].

4. Non-verbal behavior analysis unit

4.1. Selection of parts with data glove

In this system, the analysis of spoken language and that of behavior are performed in parallel. This makes it possible for a user to specify the mechanical part to be manipulated or selected by pointing action, or to grasp and move it by hand using spoken language.

At present, the analysis of user's behavior is limited to only the hand movement. Using three-dimensional position sensor added to the user's writs, the quantity of the translation and rotation are measured from the wrist. The attitude of the hand is detected using a data glove.

The system permits the user to specify an object by pointing with a forefinger. The state of the hand is judged referring to the values of joint angles.

When a data glove is pointing at some objects as shown in the Figure 6, the system judges that the parts are to be selected that are included in a cone emanating from the finger-tip and that are intersecting the conic beam

When the palm of the data glove is going to be closed to grasp the part as shown in the Figure 7, if a user has no objects in the data glove and if the bounding box of the data glove and the bounding box of the parts interfere each other, the system decides that the user has grasped the part and changes the color again.

4.2. The analysis of the hand movement using a three-dimensional position sensor

When an object is selected using a data glove, the user may move his/her hand with the forefinger pointing out. Of

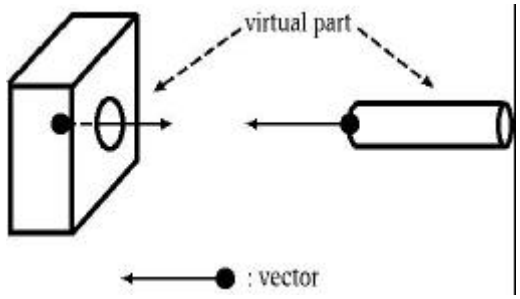


Figure5. A virtual part given an arrow

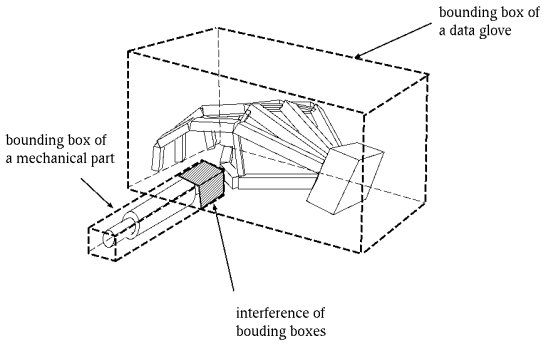


Figure 7. Selection of parts by grasping operation

course a forefinger may be bent when the user intends to point at nothing. In the former case, the cone emanating from the forefinger may interfere in several objects during the hand displacement. It is, however, difficult to find the object that the user aimed at from the interfered objects.

As a result of analyzing the behavior of a man, when the man points at an object with a forefinger, the hand generally stops with the forefinger pointing at the object for a while.

The movement of the hand measured with a three-dimensional position sensor when a man is going to point at an object is shown in the Figure 8.

It is understood that the pointing action corresponds to the portion (a) in the Figure, and that the data from the three-dimensional position sensor are comparatively stable for the moment. So on finding that the movement of the hand stops, the procedure mentioned in 4.1 is made active.

5. Spoken language processing unit

5.1. Natural language processing

A spoken language input from a user (Japanese) is converted into a character string by JULIAN. Next, a natural language processing program will analyze the character

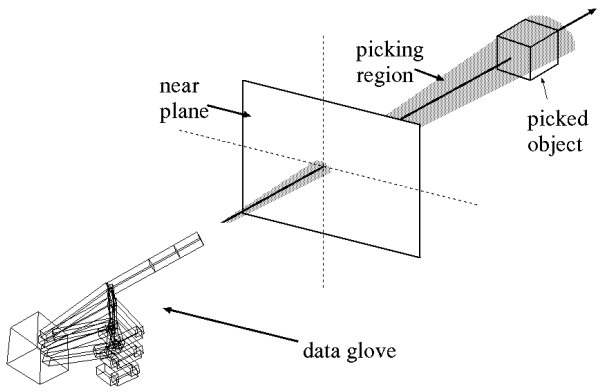


Figure 6. Selection of part by pointing action

string through the speech recognition and the semantics of the voice is extracted. A user must register into JULLIAN the words and syntax rules used in the speech recognition as described in 2.2. The dictionary made at that time is also available to the language processing.

We show an example of the word dictionary and syntax dictionary in Figure 9 and Figure 10.

The syntax rule is registered assuming the categories registered in the dictionary to be non-terminal symbols.

Semantic analysis is done in top-down fashion. When a sentence “Assemble the worm shaft.(ウォームシャフトを組み立てろ．)” is input, the input sentence is matched to the syntax “OBJ WO OPV_A AUX_A” in the syntax dictionary, and a category shown in the Figure 11 is obtained.

Because a category is registered corresponding to a function of a word, semantics of the word becomes possible.

At this time whether the content of the sentence can be handled with the system or not is judged. To increase a number of sentences to be understood, you have only to add words belonging to a category or categories and syntax rules. Inversion expression and more than one expression can be also accepted. The second rule in syntax rules shown in the Figure 10 is the inversion form of the first syntax rule. The flow of the process is shown in Figure 11.

5.2. Constructing the contents of dialog

An analysis result provided with the natural language processing exploits the knowledge of assembly, and is stored in a list called a contents list. When “Assemble the worm shaft.(ウォームシャフトを組み立てろ．)” is a recognized character string, a contents list as shown in the Figure 12 is made.

Key words corresponding to the contents of the sentence

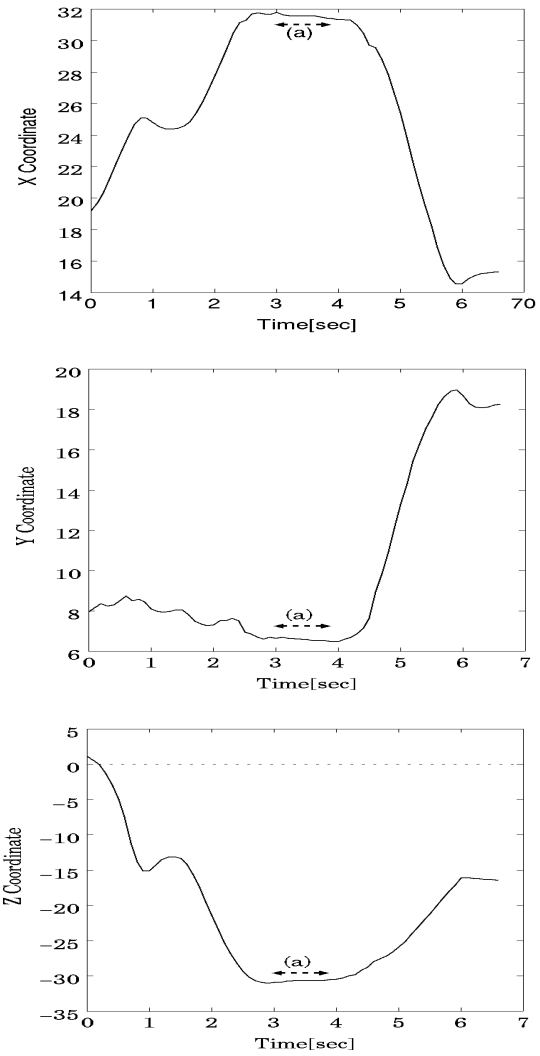


Figure 8. Analysis of the hand movement using a 3-dimensional position sensor

are stored in the first line of the contents list. The system distinguishes the contents of the utterance with the key words. If some assembly operation is necessary, a method realizing the operation is put in the second line, and the object is entered in the 3rd and the 4th line. Because two parts are mainly selected as objects of one operation, the 3rd and the 4th line are prepared. Voice information from a user is transmitted to the avatar unit in the form of the content list.

5.3. Flow of process

We describe the flow of process of spoken language in the following. At first, a user issues an inquiry or command to the system using a spoken language. Next, natural language processing analyzes the spoken language, and if the system is able to accept the contents, a contents list is made. Otherwise, the user must repeat the voice input.

The content list of the conversation is estimated after being communicated to the avatar unit.

```
##An object of operation → Semantics category
%OBJ → ory
name
グリース The word to recognized
ウォームシャフト
  :

##Method of operation (～ます)
%OPV_A
組み立て
取り付け
  :

## The end of a verb (imperative)
%AUX_A
ろ
なさい
  :

##Particle (を)
%WO
を
```

Figure 9. A part of dictionary

```
OBJ WO OPV_A
OPV_A AUX_A OBJ
```

Figure 10. A part of syntax dictionary

6. Instructor (avatar) unit

In this chapter, we explain the process performed in an instructor unit. The nonverbal information and the spoken language information from a user are respectively processed in the nonverbal behavior processing unit and the spoken language processing unit and their results are communicated to an instructor unit.

An instructor unit evaluates the information and makes the appropriate response to a user based on the factual knowledge of virtual parts described in a system.

On the instruction of parts operation, it is important to have a user operate a mechanical part with a data glove, but we believe that he will understand how to manipulate the part if he sees someone operating the part. So in this system, we prepare the following mode for responding to a question or a command from a user. In the mode, an avatar shows a user how to operate a virtual part with his hands explaining the operation in a spoken language.

This chapter explains the process of behavior generation of an avatar in detail.

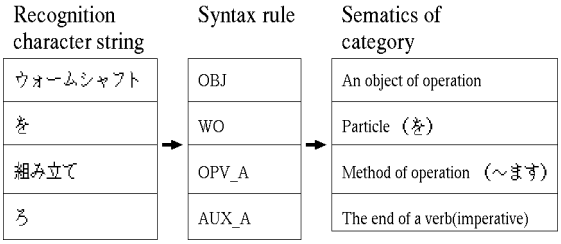


Figure 11. A result of language processing

Order	Operation
Operation	assemble
Object1	worm shaft
Object2	Nothing

Figure 12. Contents list

6.1. Dialog engine

A dialog engine shown below is installed into the instructor unit to make response for a user.

When an operation command is given from a user, the system matches the content list (5.2) to the operation prescribed in the AND/OR graph (3.3), then a response is made. If the operation command fits the operation in the AND/OR graph, the operation and explanation are performed by an avatar. When the instruction is wrong, a warning is issued. On the contrary, an avatar refers to the AND/OR graph to generate a command or question and the current state of the world. As the system (an avatar) knows the current state of the world, it is able to generate both the relevant command and erroneous command referring to the AND/OR graphs. In almost cases, a relevant command will be entrusted to a user.

When he cannot show a correct answer, an avatar will show the answer to him by moving parts. If he hesitates without starting action, an avatar asks if a user understand what he should operate. At first, he is asked to tell two-part names and to indicate them. If he cannot respond, the avatar will show the answer in place of him. If he is going to grasp a wrong part, the avatar will show the right one. If he is going to move the part toward wrong direction, it is prohibit and the avatar will show the correct movement by his hand in the same way as the case a user takes a leadership.

Voice output is generated with the ViaVoice of IBM, but it simply translates a sentence generated with the answer generating routine into Japanese emotional voice output.

6.2. Correspondence between verbal information and a nonverbal one

We have already described that in this system we prepared a mode to designate mechanical parts by combining a directive with a pointing action.

Correspondence between a directive and a pointing action is necessary for a system to understand the contents which the directive shows. This process is performed in the avatar unit in which a verbal information and nonverbal one are collected from a user.

For example, when “install the worm shaft here” is input, the object or the place corresponding to the phrase “the worm shaft” and the word “here” are found from the pointing actions, respectively. In this case, even if there were several parts corresponding to “the worm shaft”, the one belonging to the class of a worm shaft is put into the candidate set. Nonetheless, when two or more candidates are left, the one with the size or the structural characteristic making the operation specified possible is selected. Nevertheless, if a unique object cannot be determined, the system must ask a question to the user to make clear an object to be selected.

Note here that there is a problem. The analysis of user’s action is taken place in real time because it is not measured with a vision system but with magnetic sensors. On the other hand, as the analysis of utterance is prolonged until it will terminate, it is hard for the system to know the word uttered as soon as corresponding action is analyzed. When actions that are recognized as pointing action are observed several times and several demonstrative pronouns or the definite names appear in the corresponding utterance, to find the correspondence between the actions and pronouns/nouns is difficult. If the numbers of their appearance are equal, then they correspond in the order of appearance. At first, the system solved the problem based on the order of appearance, but at present the correspondence is solved based on the time of appearance.

6.3. Behavior generation of an avatar

An avatar has three joints in his arm and 14 joints in his hand same as shown in the Figure 13.

The behavior of an avatar is decided by assigning respective values to the position and rotation of each joint. The values given to the position and rotation of each joint are constrained as the attitude of an avatar cannot deviate from the human attitude.

In this system, the avatar is permitted to do nonverbal behavior such as grasping and pointing action in a virtual environment. As it is difficulty to compute all values of position and rotation of 14 joints of his hand in case of both behavior, they are acquired with the motion capture

method using a data glove beforehand.

Next we explain how to determine the attitude of his arm. Here presume that the position of his shoulder is fixed in the pointing or grasping action. Then everything to be done is to determine the values of the position and rotation of remaining wrist, elbow and shoulder.

In case of the pointing behavior, the center of gravity of the part an avatar is going to point is first retrieved. Next, the values of the position and rotation of a wrist are determined to enable him to point to this centroid position. Values of the position and rotation of the remaining elbow and shoulder can be obtained by the inverse kinematics.

As how to grasp depends on the shape of the object to be grasped, it is difficult to decide values of the position and rotation of a wrist by computation. Consequently relative position between the part to be grasped and a wrist must be registered beforehand. Values of the position and rotation of the remaining elbow and shoulder can be obtained in the same way as in the pointing action.

A series of attitudes of an avatar from the initial state to the end state can be got by a linear interpolation of both situations. When an avatar can neither point to nor grasp a part from his current position, he has to moves to the new position that makes him do the behavior.

Figures 14 and 15 show the situation an avatar performed the installation operation of the worm shaft.

6.4. Synchronization of behavior and spoken language by avatar

Problem is the synchronization of voice and behavior. When an avatar mates a part A and B, the way of operation must be explained with voice while he is performing the operation.

As an example, consider the following case in which an avatar shows a sequence of operations saying that you should insert this part A into the hole of the part B. After first the avatar extends his arm to an object and grasps it, he will move it to an approach point of an operation. At that occasion, it is assumed that the reference of the part name is finished at the same time as he grasps the part.

When a necessary preparation is successfully done, the avatar explains how to mate them while operating them from approach points. Here, it is assumed that the approach points are set at the positions shown in the Figure 14. After all, the following will be his behavior. He will grasp a part uttering a phrase like “the part A”, he will move it to an approach point afterwards. The similar operation is performed on the part B. Note that he will move a part to an approach point saying nothing. He will say that “in this way a part A is inserted into B” operating the part. When an operation consists of several secondary operations, each secondary operation and a corresponding explanation are synchronized in the similar method as that mentioned above.

The ViaVoice of IBM is used to have an avatar speak the content of explanation. The software gives us the time needed to utter a phrase consisting of N characters. Of course the time needed for the utterance is also controllable, and the start and the completion time of an utterance are also controllable.

On the other hand the time needed to draw each frame of an avatar’s behavior is depended on the complexity of background (Context) in drawing and a power of a computer used.

Assume here that it takes T to utter a series of phrases and that a graphic generation of M (M=M1+M2) frames must be finished at the end of the utterance. M1 is a frame number to be drawn by directly before the time he grasps the object and it is determined in the following way. M2 means images from the grasping point to the approach point. Interpolation of images are performed using the inverse kinematics so that his hand moves smoothly along a line connecting an initial point and an end point. Let assume the background or context does not change suddenly while an avatar utters a given phrase.

And if the time needed for drawing an initial frame is t, then the time necessary for drawing of M frames becomes tM. (Of course this graphic generation is invisible to users).

In the case of T>tM1, give the sleep time(T/t-M1) after every frame generation. In the case of T<tM1, start the speech synthesis at the time(tM1-T) after graphic generation started. As the utterance is finished almost simultaneously at the time he grasps an object, the system has

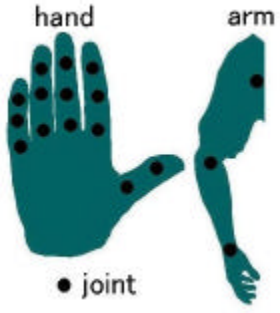


Figure13. Joint of hand and arm

Figure14. Grasp of virtual parts

Figure15. Installation of a worm

only to drawn the remaining M2 image in the same rate as M1 afterwards. And the speech synthesizer is notified of the completion of a graphic generation.

In this way, by dividing one operation into two steps, a grasp and operation itself, voice and a graphic generation are synchronized. There are some cases that an operator does not require another part to be operated. In such cases the object of the operation is considered to be just one.

7. CONCLUSION

In this research the training system, which instructs assembly / disassembly of mechanical parts to a user was developed. A bi-directional interface system is realized that permits a user and the system to communicate each other using verbal and nonverbal information.

Not only the synchronized recognition of voice and behavior of a user, but also the synchronization mechanism of the speech synthesis and the behavior generation of an avatar were stated clearly.

You may feel a few differences between the impression from the avatar and the sense received from a human being.

When an avatar utters a same word again and again, the tone should be changed. For example, even if a user repeated the same mistake, an avatar just utters the same warning in the same tone. Functions that make his tone and expression more strictly are necessary in order to give many better effects to a user.

The current system cannot prohibit any erroneous operation of a user physically. By replacing the hand of a user with PHANToM which is the haptic interface and restricting the movement of PHANToM, operation errors can be prohibited.

It is evaluated with various situations, and technique proposed with a research now is improved.

We will continue our effort aiming at the construction of more natural man machine interface by introducing new frames and evaluating them in various situations.

References

[1] Norihiro Abe and Saburo Tsuji "A consulting system which detects and undoes erroneous operations by novices" Proc. of SPIE, pp.352358, (10 1986)

[2] Norihiro Abe, Tomohiro Amano, Kazuaki Tanaka, J.Y.Zheng, Shoujie He, and Hirokazu Taki "A Training System for Detecting Novice's Erroneous Operation in Repairing Virtual Machines" International Conference on Virtual Reality and Tele-Existence(ICAT), pp.224229,(1997)

[3] Norihiro Abe, J.Y.Zheng, Kazuaki Tanaka and Hirokazu Taki "A training System using Virtual Machines for Teaching Assembling/ Disassembling Operations to Novices" International Conference on System, Man and Cybernetics,pp.20962101 (1996)

[4] J, Wernecke "The Inventor Mentor", Addison Wesley Publishing Company (1994)

[5] Tomoaki Ozaki, Kazuaki Tnaka, Norihiro Abe, Hirokazu Taki"Verbal/Nonverbal communication in Virtual Environment", International Conference on Systems, Man, and Cybernetics, 1999,to appear