VOICE DATA ENTRY IN NISTARS WAREHOUSES

James M. Stokes

CHI Systems Technical Report 921207.9203
28 January 1993

CONTRACT N00600-92-C-3043
CDRL A002

CHI Systems, Inc.
Gwynedd Plaza III
Bethlehem Pike at Sheble Lane
Spring House, PA 19477

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate of Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE 28 January 1993	3. REPORT TYPE AND DATES COVERED Final Technical Report 7/92 - 12/92

4. TITLE AND SUBTITLE

Voice Data Entry in NISTARS Warehouses

5. FUNDING NUMBERS

N00600-92-C-3034

6. AUTHOR(S)

Mr. James M. Stokes

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

CHI Systems, Inc.
Gwynedd Plaza III
Spring House, PA 19477

8. PERFORMING ORGANIZATION REPORT NUMBER

921207.9203

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Commander
Naval Supply Systems Command
SUP 4233D
Washington, DC 20376-5000

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; Distribution unlimited

12b. DISTRIBUTION CODE

A

13. ABSTRACT *(Maximum 200 words)*

The reported effort was undertaken to determine the feasibility of implementing an effective voice data entry capability for NISTARS warehouses. The Fixed Carrousel Work Station was determined to offer the best opportunity for voice enhancement, based on the degree to which data entry activities and material handling activities are interleaved at that station. A voice-interactive interface was designed which allows the warehouse worker to attend to the necessary material handling while simultaneously performing data entry. In order to implement this interface within existing NISTARS hardware and software constraints, a voice "server" strategy was developed. This open systems approach shields NISTARS from potential changes in voice technology and provides a mechanism for integrating voice in the RF-linked hand-held stations in the future. When all work activities are considered, replacing the existing manual/visual computer interface with interactive voice will result in a time savings of about ten percent. Given a work station utilization of slightly over fifty percent, this level of savings is sufficient to pay for the required equipment acquisitions within two years. Once implemented, the voice interactive system reported here could improve worker productivity and reduce response time in providing materials and services to the fleet.

14. SUBJECT TERMS

Voice recognition, speech technology, computer-human interaction, data entry

15. NUMBER OF PAGES

40

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT U	18. SECURITY CLASSIFICATION OF THIS PAGE U	19. SECURITY CLASSIFICATION OF ABSTRACT U	20. LIMITATION OF ABSTRACT

N00600-92-C-3043

TABLE OF CONTENTS

1. Introduction and Overview 1

2. Candidate Workstation Selection 2

2.1 Advantages of Voice Interface Technology 2

2.2 NISTARS Work Station Overview 3

2.3 Conflicts Between Task and Interface Activities 4

3. Carrousel Voice Interface Design 7

3.1 Visual vs Auditory Output 7

3.2 Restricting the Scope of Voice Utilization 7

3.3 Location Verification by Voice 8

3.4 Concurrence in Task and Interface Activities 9

3.5 Vocabulary and Dialog Specifications 11

3.5.1 Dialog Overview 11

3.5.1 User Requests for Voice Output 16

3.5.1 Error Handling 18

3.5.1 Voice Input Vocabulary 19

3.6 Template Training 20

4. Voice Interface Implementation Plan 21

4.1 Integration Options 21

4.2 Voice Server Specifications 25

4.3 Task Breakdown 25

5. Cost Benefit Assessment 28

5.1 Overview 28

5.2 Expected Benefits 29

5.3 Expected Costs 31

5.4 Recommendations 32

References 34

Appendix: DDD Jacksonville Employee Statistics 35

1. INTRODUCTION AND OVERVIEW

The implementation of an effective voice data entry capability for NISTARS warehouses could improve worker productivity and reduce response time in providing materials and services to the fleet. The Phase I SBIR work reported here was undertaken in order to determine the feasibility of such an implementation. A voice-based interface was designed for the NISTARS work station which showed the greatest potential for enhancement through the application of voice technology. Following a review of existing NISTARS hardware and software configurations, an implementation strategy for incorporation of voice and support of the interface was developed. Based on the interface design and implementation plan, benefits and costs were assessed to determine feasibility.

The Fixed Carrousel Work Station was determined to offer the best opportunity for voice enhancement, based on the degree to which data entry activities and material handling activities are interleaved at that station. A voice-interactive interface was designed which allows the warehouse worker to attend to the necessary material handling activities while simultaneously performing data entry. A seventeen word recognition vocabulary for voice input, combined with a series of voice playback messages, forms the basis for a human-machine voice dialog which eliminates the worker's need to return repeatedly to the computer station for keyboard data entry and visual data retrieval. In order to implement this interface within the existing NISTARS hardware and software constraints, a voice "server" strategy was developed. A single server unit, constructed from off-the-shelf components, will provide voice interaction capabilities for multiple carrousel work stations. This open systems approach allows the specialized, low-level aspects of voice interaction to be hidden from the work station interface software. The larger NISTARS system is buffered from changes in voice technology and provided with a mechanism for integrating voice in the RF linked hand-held stations in the future.

The time savings offered by interactive vóice over the existing manual/visual computer interface is about fifty percent, but this advantage is reduced to about ten precent when all other work activities are considered. Based on the interface design and a familiarity with the characteristics of automated voice, the series of movements, gestures and utterances associated with each of the interface versions was performed and timed. The time difference for the data entry activity was found to be eight to nine seconds in a task requiring one to one and a half minutes. Given a work station utilization of slightly over fifty percent, this level of savings is sufficient to pay for the required equipment acquisitions within two years. Once developed, the voice server can also be utilized for future voice implementation of hand-held stations, phasing voice into high payoff areas such as cold storage. The ability to pick orders more quickly is itself a benefit of major significance since it goes directly to the need to reduce response time in providing materials and services to the fleet.

2. CANDIDATE WORKSTATION SELECTION

2.1 Advantages of Voice Interface Technology

The application of voice technology to NISTARS workstations on the warehouse floor is consistent with the series of options available for the application of voice generally. The niches where voice technology confers an advantage over other interface technologies are now well known. Three principal areas have proven to be central to the application of automated voice. First, there is the possibility that a target user population is unable to use, or is averse to using, other interface techniques such as keyboard entry. Second, the possibility exists that data entry is actually faster using voice in place of other techniques. Third, there is also the possibility that data can be entered by voice in parallel with other activities, an advantageous characteristic in a number of contexts. All three areas must be examined in order to determine the appropriate application of voice technology to NISTARS.

The first type of advantage for voice is related to abilities or attitudes with reference to other technologies. In cases of handicapped users, such as various forms of paralysis, voice is one of the few interface devices available to the user. Voice input has clearly established itself in this application area. Whether or not handicapped applications are of interest in the NISTARS context, the same handicaps which prevent users from functioning with keyboards also prevent individuals from performing the many manual tasks which are the central activities of the warehouse worker. Voice technology cannot therefore confer any advantage due to limitations on user's interface capabilities although there are undoubtedly tasks outside of those on the warehouse floor which could be made accessible to the handicapped with the incorporation of voice. These areas are not included in the scope of the effort reported here. Another advantage which focuses on the negative aspects of other technology relates to the perceived low status of key entry and control. Voice applications have been developed for users who need to communicate with computers but virtually refuse to use keyboards. Building medical record keeping software for use by doctors is of current interest in this area. Numerous other attempts have been made to sell PC-based voice systems for use by middle and high level managers in business. Like the handicapped situation, there are undoubtedly warehouse applications that fall into this domain but are likewise outside the focus of this effort.

The voice typewriter, the Holy Grail of voice recognition, is still not a reality but epitomizes the dream of efficient data entry by voice. If available, only the fastest stenographic typists could hope to enter text as quickly as machine transcription by automated voice. The large vocabulary recognizers which have become available in the last ten years have brought applications of a similar sort into being. Their significance depends on their ability to provide a large number of direct entry items which would be difficult to match in a single set of hard keys. In the voice typewriter, each of thousands of input words would act as a dedicated button for a specific series of entry characters - the equivalent typed word. At some level of required vocabulary size, word entry becomes more efficient than menu and soft-key strategies. In this sense, voice input can be more "natural", in that it allows direct input of items of interest without composing them of lower level elements or operations. Direct entry of simple letters and numbers by voice, however, is still a slower process than entering the same data by keyboard or keypad. This is because it takes longer to utter even a short word than to press a key. For numeric entry it will probably remain true even after the ultimate advent of voice typewriters. A hand poised over a numeric pad will still be faster than calling out numbers and, even with the dangers of carpel-tunnel syndrome, the use of voice input would probably prove less user-friendly than key input for large amounts of data. The type of complex filling-in of forms, such as in medical test evaluation, for which the current large vocabulary recognizers seem to offer some speed advantage, do not occur in NISTARS. Most data entry here is simple numerics with interspersed, isolated response or command keystrokes. A direct run-off between between voice input and key input for these items would prove key input the winner.

The hands busy context provides existing voice technology with its most commonly occurring and most powerful advantage. In its most extreme case, the user is completely prevented from utilizing manual interface techniques. For example, in the case of a fighter pilot, both hands may be occupied when it becomes necessary to tune a radio. By off-loading the latter function to voice, the two tasks of flying and tuning can be performed in parallel. In less extreme cases, such as analysis of x-rays, quality inspection and so on, it is efficient to continue visual examination of some object while entering results. While some types of key input could be performed in parallel with this type of eyes busy activity, it is seldom convenient due to changing body position or actual use of hands in the inspection process. The most common, although perhaps the least obvious, form of the "hands busy" situation is mutual disruption between physical, non-data-entry task activities and the data entry itself. Whenever the inter-leaving of data entry activities with other task activities causes the worker to repeatedly change position or posture, the two activities can be said to be mutually disruptive. Mutual disruption of this type is a characteristic shared to a greater or lesser degree by all the NISTARS workstations on the warehouse floor. The choice of a candidate workstation for voice interfacing is principally a matter of determining where the possibility of avoiding mutual disruption suggests the greatest advantage for voice over key input.

2.2 NISTARS Work Station Overview

Although other work stations exist within the NISTARS system, four stations situated on the warehouse floor were the focus of the effort reported here: Induction, Fixed Carrousel, Hand-Held, and Consolidation and Packing work stations were examined as potential candidates for voice interfacing. The extent to which the performance of data entry tasks at these stations can be performed in parallel with other activities and the extent to which such parallelism can enhance worker production levels is here summarized. This material forms the basis for the selection of a single candidate work station, a detailed examination of which follows in the next section.

The Consolidation and Packing Work Station offers little opportunity for voice interface integration. Normal processing requires no more than a <Task Complete> key entry to the computer. Even exception processing seldom requires more interaction than the entry of a yes/no answer or a single number. There is no consistent breakup of the flow of non-interface activities at the work station as a result of these occasional entries. Off-loading the key input to voice would have some minor impact on the time required to perform some exception handling, but overall the effect would be minimal.

Induction presents a similar picture. Although there is more data entry in normal processing at this work station than for Cosolidation and Packing, it is not particularly disruptive to non-interface activities. Under Direct Enhanced ABE Receiving (DEAR), the most common type of receiving now performed, three data items are typed in during normal processing - signal code, fund code, and quantity. Both signal code and fund code are read off the document which accompanied the shipment. This requires that the worker move to the document whether or not the codes are entered at a keyboard. If the worker is already positioned in front of the terminal with the document at hand, using voice input might actually slow down this portion of the induction process. If only entering quantity can be implemented to some advantage using voice recognition, the overall effect would probably be minimal, as in the Consolidation and Packing case. In addition, once voice recognition of digits was implemented for the entry of quantity, it is possible that workers would begin to use it for signal and fund codes and actually slightly degrade overall performance.

Carrousel and Hand-Held work stations support store, issue, and inventory functions utilizing very similar interface and activity sequences. Although all of these functions offer considerable promise, the issue or pick function provides the clearest example of a process which interleaves data entry with physical activities. There is an obvious opportunity here to enhance production by allowing the worker to perform data entry in parallel with other required activities. As the Carrousel and Hand-Held work stations are very similar in this respect, the selection of a final candidate must be determined by overall impact. NISTARS Type 2 installations were designed with carrousels as the main fast pick station. The small, high volume items such as nuts and bolts stored on the carrousels give way to items such as fittings and pipes in the binnable areas and turbine parts in rackable storage. As handling larger objects is generally more time consuming it is likely that data entry activities take a proportionally smaller amount of time when picking larger objects. Greater percentage reductions in task times are therefore likely at the small item, high volume end of the range. The Carrousel work station has therefore been selected as the best candidate for initial voice interface development. The resulting increase in production per unit of time should exceed that for any other NISTARS work station so implemented.

2.3 Conflicts Between Task and Interface Activities

Figure 2-1 presents an analysis of the normal carrousel pick procedure, as it is currently performed, with time moving forward from top to bottom of the chart. Task activity refers to what, in general, the worker is attempting to do to accomplish the pick task. Interface activity refers to actions taken to retrieve information from or enter information into the computer. Physical activity refers to the manipulation of non-interface objects and movements intended to change the worker's position relative to computer, carrousel, work table or bin for completed work. It is obvious from the repetition of the word "turn" in the physical activity column that the warehouse worker is constantly moving back and forth between several positions or postures, frequently as a result of the requirement for interaction with the computer.

Task Activity	Interface Activity	Physical Activity
Determine shelf of interest	Read text for shelf location	Approach computer
Find and verify shelf of interest	Enter location label	Take wand from computer Turn, approach carrousel Gesture with wand
Determine box of interest	Read graphic for box location	Turn, approach computer Place wand back on computer
Find box of interest		Turn to carrousel
Examine box contents for NSN		Approach carrousel, look inside box
Verify NSN correct	Type in last two NSN digits	Turn, approach computer Type on keyboard
Determine pick quantity	Read text for quantity and limits	
Retrieve box		Turn, remove box from carrousel Turn, carry box to work table
Unpack box contents (if necessary)		Open and empty bag(s) or box(es) (if necessary)
Count out order		Count out required quantity of items
Bag picked items		Place picked items in new bag and seal
Identify issue-bag for system	Enter PIN on bag of items	Turn, carry new bag to computer Gesture with wand
Enter and verify pick count	Type in pick count	Type on keyboard
Set aside completed issue-bag		Turn, carry new bag to cart of completed work
Repackage box contents (if necessary)		Turn, approach work table Place remaining items in original bag or box (if necessary)
Return box to carrousel shelf		Turn, carry box from work table and place back on carrousel
Terminate task	Press <Task Complete> key	Turn to computer

Figure 2-1 Current Carrousel Pick Procedure

Exclusive of the data exchange between worker and computer, the pick task breaks down into the following steps:

- find container
- retrieve container
- select items
- pack items
- set aside items
- replace container

With the ideal interface, the warehouse worker would be able to perform these steps without ever disrupting the activity flow to access the computer. Unfortunately the following five input actions must be taken by the worker during the pick task:

1 verify location to system
2 verify container contents to system
3 inform system of PIN
4 inform system of pick count
5 inform system that task is completed

Actions 2, 4 and 5 are currently performed by keyboard entry and can readily be replaced by voice recognition done in parallel with the task steps just noted. Actions 1 and 3 are currently performed by means of the bar code wand. In order to eliminate the need to return to the computer to perform these actions, workers must either have the wand located where it is used, carry the wand with them (by Velcro attachment to clothing, for example), or be allowed to substitute a voice-based procedure. It should be noted here that there are also several read operations shown in Figure 2-1. These will be discussed in the interface design section which follows.

If it is possible to treat all of the interface activity shown in the figure as operations performed in parallel to more basic pick activities, the time savings could be significant. It should be kept in mind that once these activities no longer require access to the computer, the need to perform a good deal of the physical activity shown will be eliminated as well. Although the carrousel store and inventory operations are less complicated examples of data entry and physical task activity being mutually disruptive, they are similar in form and offer some opportunity to trim task time requirements by applying voice technology.

3. CARROUSEL VOICE INTERFACE DESIGN

3.1 Visual vs Auditory Output

As stated in the Phase I proposal, the effort here described is an attempt to apply voice recognition technology to NISTARS, in the context of visual as well as auditory system output. It is clear from the argument made for the carrousel work station above, that the existing visual feedback channel (the PC monitor) is unacceptable. If voice recognition replaces keyboard input, the warehouse worker need not return to the PC in order to enter data but may enter data while continuing to perform other task activities. If, however, the worker still needs to return to the PC to read data items from the computer screen, often on a one-for-one basis with the data entry items which have just been divorced from the PC location, the gains realized by applying voice will be lost. It is possible to provide small headset mounted displays which could free the worker from the PC monitor, but such technology is a good deal more uncomfortable than a simple microphone headset. It is also difficult for users to adjust to such displays when looking at many objects at varying distances causing frequent changes in the distance at which the eyes are focused. No viable visual feedback mechanism has been identified in the course of this project. However, since the playback of voice messages is provided on most recognition boards, voice feedback hardware is already available once the commitment is made to voice recognition. Voice output has the added advantage that its combination with voice input is a natural pairing, creating a human-machine voice dialog. The use of voice output can help remind the user that this is a special type of voice communication where a new type of voice control must be exercised to assure recognition on the part of the (machine) listener. The design described in the sections which follow does not include the development of a specialized visual feedback variant, but rather assumes that the existing carrousel work station display will continue to function virtually unchanged.

3.2 Restricting the Scope of Voice Utilization

With both voice recognition and voice playback available, it is possible to develop a system which is totally based on voice I/O -- a continuous voice dialog between worker and computer. However, this approach would not necessarily produce an interface which was more efficient than the one which currently exists. We want to exploit the ability to perform task components concurrently using voice while not extending voice to data entry or control areas when it does not produce parallel activities or provide a superior mechanism by itself. If all the human-machine interaction for the Carrousel work station functions should not be implemented by means of voice technology, the question remains as to where to locate the dividing line between voice and non-voice. Researchers have pointed out that speech must be assigned in a consistent way to one task component or to a small sub-set of commands or data (Jones, Hapeshi & Frankish, 1989). Since the normal pattern of pick (or other function) operations is to consistently interleave interface and non-interface activities such that those operations offer the main target for time savings, it is proposed here that exception handling be excluded from the voice interface. The warehouse worker will be aware that the task components routinely accomplished by means of voice form the limits for the use of the new technology. With this approach the vocabulary can be kept small and easily remembered and areas which are difficult to handle by voice, such as the visuals in the Carrousel Configuration Display, will be avoided. Since voice technology is new to NISTARS personnel, it is also important that the interface be kept simple to improve user acceptance. Although it will be necessary to provide voice messages to indicate the need for exception processing, workers are not likely to mistakenly attempt to use voice input in the absence of voice prompts. Existing procedures will be used for handling exceptions, based on the existing work station interface.

3.3 Location Verification by Voice

The fact that the bar code wand is attached to the PC forces the user to move to the PC to retrieve the wand and move back to the PC to return it when the wanding operation is completed. This is exactly the type of physical activity we would like to eliminate through the application of voice technology. In the case of the PIN entry during carrousel picks, voice cannot easily provide a more efficient solution. The number is long and the time taken to enter it by voice would be no less than that taken to perform the physical movements necessary to wand it. Since the number is entered near the end of the pick operation, it cannot easily be done in parallel by voice during other activities. That is, the worker at this point in the procedure is ready to set aside the selected items but cannot do so until the PIN is entered because the number would have to be read off the prepared bag. Unlike the PIN entry, the location verification wand operation at the beginning of the carrousel pick (and at the beginning of other carrousel functions) can be replaced by a voice operation.

The existing system requires the worker to wand the bar code label of the shelf of interest, as determined by examination of the computer display. This operation verifies that the correct carrousel, stack, and shelf have been identified by the worker, at least at the instant when the wanding occurred. Unfortunately, the worker must turn around, return the wand to the PC mount, and then turn back to access the container of interest. Upon return to the stack, it is quite possible for the worker to go to a container which is a shelf above or below the one wanded. The entry of the NSN digits is not a particularly good control over this type of error. Firstly, the system does not guarantee uniqueness of the digit pair within the stack or row, although it could do so during the stow operation. Secondly, the NSN is not a simple sequence number, but rather contains type codings so that some final digit pairs may occur quite frequently. The chances for pick errors in this system seem relatively high but would require analysis of mistakes in shipped orders and failures to pick in order to assess actual rates. Voice options would appear to offer control over workers access of the correct location with dependability at least as good as the existing system.

Verification of location and container contents can both be performed while the worker is facing the carrousel. For example:

1 system says “carrousel B, stack twelve”
2 user says “OK”
3 system says “row C box two”
4 user says “three eight”

The worker is first told the stack location by voice. When the worker has visually verified the stack and entered confirmation by voice, the system calls out the shelf and box location. The worker responds with the last two digits of the NSN found in the box and proceeds. Unfortunately, with this scheme, location verification is completely under control of the worker and can effectively be ignored. The worker could simply respond “OK” without ever checking the stack markings. An alternate scheme, similar to the existing procedure for contents verification, is recommended as part of the interface design:

1 system says “carrousel B”
2 user says “one two”
3 system says “row C box two”
4 user says “three eight”

In Step 1 the system tells the worker which carrousel contains the items to be selected and the worker responds by entering the number of the stack which has been positioned for the pick operation. Steps 3 and 4 are identical to the previous example. At no point does the worker turn away from the carrousel and, therefore, there is little chance to access a location which does not correspond to the verification data entered. It is certainly possible that warehouse workers sometimes examine the contents of a box, enter the NSN digits, and then return to a neighboring box to acquire the items to be issued. The occurrence of this type of access error should also be eliminated by the above dialog structure since the worker will enter the NSN digits while removing the box from the shelf.

3.4 Concurrence in Task and Interface Activities

Figure 3-1 presents an analysis of the carrousel pick procedure, as it will be performed when the voice interface is in place. As explained above, only normal processing is presented, utilizing voice output and voice verification of location. Comparison with the analysis in Figure 2-1 immediately shows a reduction in the number of steps require to complete the pick task. A more detailed examination will also show an increase in the concurrent performance of interface and non-interface tasks. For example, the only concurrence of this type in the existing system is the ability to “Read graphic for box location” at the same time as “Place wand back on computer”. In the voice system, the ability to “Listen to quantity” at the same time as “Remove box from carrousel” is one of many similar opportunities to perform two actions in parallel.

Task Activity	Interface Activity	Physical Activity
Determine shelf of interest	Listen to shelf location	Approach carrousel
Verify location found	Say two digit stack ID Listen to box location	
Examine box contents for NSN and verify NSN correct	Say last two NSN digits followed by "enter"	Look inside box
Determine pick quantity and retrieve box	Listen to quantity	Remove box from carrousel and turn, carry to work table
Unpack box contents (if necessary)		Open and empty bag(s) or box(es) (if necessary)
Count out order		Count out required quantity of items
Bag picked items		Place picked items in new bag and seal
Identify issue-bag for system	Enter PIN on bag of items	Turn, carry new bag to computer Gesture with wand
Enter and verify pick count and set aside completed issue-bag	Say item count followed by "enter", listen for system "OK"	Turn, carry new bag to cart of completed work
Repackage box contents (if necessary)		Turn, approach work table Place remaining items in original bag or box (if necessary)
Return box to carrousel shelf	Say "completed"	Turn, carry box from work table and place back on carousel

Figure 3-1 Voice-Based Carrousel Pick Procedure

The reduction in required task time which results from the increased concurrence can best be summarized by examining the changing physical position of the worker during the course of the pick task. Under the current interface, physical attention (hands on) at different locations at the work station is required in the following sequence:

1 PC
2 carrousel
3 PC
4 carrousel
5 PC
6 carrousel
7 table
8 PC
9 bin
10 table
11 carrousel
12 PC

Steps 6 through 11 are required by the need to perform the non-interface tasks (described in Section 2.3 above), plus the need to wand the PIN of the issue unit. With the exception of providing the wand at the bin or table location, little can be done to streamline this portion of the pick procedure. The movement sequence represented by Steps 1 through 5, however, can be eliminated by the use of voice. Under the voice interface, physical attention at different locations is required in the following sequence:

1 carrousel
2 table
3 PC
4 bin
5 table
6 carrousel

Unlike the current procedure which begins and ends in front of the PC, the new procedure will begin and end at the carrousel. Overall, the number of position/posture changes has been reduced from eleven to five. The total time saved by converting to a voice interface includes the time previously used to make those movements or position changes which are no longer needed as well as the time previously used to perform non-parallel activities such as key entry which will now be done concurrently with some other required activity.

3.5 Vocabulary and Dialog Specifications

3.5.1 Dialog Overview

The simplest voice interface to NISTARS and many other applications would consist of a "voice keyboard", a system where each voice entry results in the entry of one or more characters as if typed at the keyboard with entry feedback appearing on the computer display. In this context, interface software would remain unchanged with the addition of voice. The application software would not even be aware of the input source, just as it currently cannot distinguish between key and wand digit entries. Unfortunately, as described above, in order for voice input to enhance system performance through activity concurrence, the user most also be given voice output to remove the constraints imposed by visual feedback. In this context, input words are no longer simply "voice buttons" but rather part of an ongoing human-machine speech dialog. Since the interface now begins to take on the appearance of a human-human verbal exchange, the similarities and differences between these two types of dialog need to be addressed. The dialog should not violate basic speech interaction characteristics in such a way that the user will find it difficult to adjust, nor should user expectations be raised to an unrealistic level regarding the possibilities of this "conversation".

One critical area where voice automation should come as close to natural speech as possible is timing. Turn-taking in human conversations is often separated by intervals of less than 50 ms (Karis & Dobroth, 1991). While computer users may be willing to accept delays in the response to keyboard entries, delayed voice responses are usually found to be annoying based on the human-human dialog model we all have internalized for speech. Not only do delayed responses slow down task performance but they also make acceptance of the technology a more difficult process. The timing of human responses to machine voice output can also cause difficulties. Most speech hardware is not capable of listening while speaking. As a result, the human speaker must wait until the machine speech is completed, if not slightly longer, before beginning a response. This characteristic can be quite annoying since the user's input will probably not be correctly recognized if the timing constraint is not observed. Another timing issue relates to the difference between discrete and continuous word recognizers. With discrete recognizers, the user must distinctly separate entry words. For example, numbers are normally spoken by running all the digit words into a single continuous utterance. For a discrete recognizer to function, small pauses must occur between the digits. Although users eventually adapt to the requirement, it can be difficult for some individuals to learn this type of voice discipline. Continuous speech, or connected word, recognizers allow the speaker to run words together as in normal conversation, although recognition accuracy may be slightly reduced as a result.

Timing issues must generally be addressed at the level of hardware and software implementation (as discussed below) and should not constrain the structure of the human-machine dialogs themselves. Of more concern for dialog construction are those areas where natural speech and automated speech diverge. The most pronounced difference between normal human conversation and automated voice is the restricted vocabulary found in automated systems. Voice data entry systems are frequently limited to a vocabulary of ten to twenty words. In this context the notion of modeling is central. Human speakers tend to model their style of speech after the style to which they are listening (Zoltan-Ford, 1984). This includes the imitation of all aspects of speech from pronunciation to vocabulary and syntax choices. In order to support the user voice discipline required for recognizer input, system voice output should avoid wordy prompts and requests and, when possible, use the same words and constructions for output that are expected for input. Although habit is ultimately the best control over input word selection by the user, a constrained output vocabulary continually reminds the user that this is not a normal conversation and, as a result, the user is not likely to use words the machine cannot recognize, or slip into different or more relaxed styles of pronunciation. Taking advantage of modeling to promote user adjustment to the needs of voice recognition is one of several reasons for providing a compact, terse output style, perhaps the most obvious being the desire to make the whole process minimally time consuming. Extended verbal prompts are also to be avoided since they quickly become annoying (Leiser, 1989). Unlike repetitious prompts on a computer screen which can simply be noted by the user without actually being read, verbal prompts take up considerable time and cannot be minimally attended to in the same way.

Figures 3-2 through 3-4 present the voice dialogs designed for normal stow, issue, and inventory procedures at the fixed carrousel stations. (Location survey is similar to inventory and is not shown in a separate figure.) Utterances of both operator and machine are kept brief, with prompts often reduced to the data which must be confirmed. Since the tasks are highly repetitive, full verbal instructions for worker actions are not appropriate. Words such as "side" are included to provide enough

Operator	Machine
	← “side B”
“one” →	
“six” →	
“enter” →	
	← “three A two”
“two” →	
“four” →	
“enter” →	
	← “ten of unit package”
wand PIN	
	← “enter count”
“one” →	
“zero” →	
“enter” →	
	← “OK”
“completed” →	
	← “side A”

Figure 3-2 Normal Carrousel Pick Dialog

Operator	Machine
	← "side B "
"one" →	
"six" →	
"enter" →	
	← "three A two"
"two" →	
"four" →	
"enter" →	
	← "OK"
wand SIN	
"completed" →	
	← "side A"

Figure 3-3 Normal Carrousel Stow Dialog

Operator	Machine
	← "side B "
"one" →	
"six" →	
"enter" →	
	← "three A two"
"two" →	
"four" →	
"enter" →	
	← "one six two three three nine"
	← "four nine seven one two four"
	← "purpose two, condition one"
	← "OK?"
"yes" →	
	← "enter count"
"one" →	
"zero" →	
"enter" →	
	← "OK"
"completed" →	
	← "side A"

Figure 3-4 Normal Carrousel Inventory Dialog

context to properly track the procedure. The use of the word "enter" to terminate the NSN digit entry is not required by the input, which is of fixed length. The usage is recommended (although the alternative should be tested during early stages of implementation) since item counts require a termination and it could be confusing to provide both terminated and non-terminated versions of number entry.

The location and contents verification steps are identical across the three procedures. All three procedures are terminated with the input of "completed" which immediately initiates the next location verification step. In other details, the procedures vary and are shown in their simplest form. If the pick procedure required the entry of a perpetual inventory count, this request would replace the "OK" response upon entry of the PIN count. If required, a remarking instruction would occur as part of the pick quantity request. A variety of messages will be required to bring the many possible exceptions to the attention of the user. Exception handling will occur through the existing keyboard/screen interface, returning to voice prompts where continued normal processing is possible. Some adjustment to these procedures may be required once an initial implementation is exercised by warehouse personnel. For example, although the verbal presentation of the complete NSN for verification in the inventory task may provide a better check than visual examination of the target number, listening to such a long string may prove tedious, or in fact, be slower than the existing procedural step. If necessary, voice output could direct the worker to compare screen and object versions of the NSN as in the current interface.

3.5.1 User Requests for Voice Output

The fact that voice is a transient communication medium can frequently make voice interfaces unfriendly. If a user is dependent on hearing a prompt or data item before proceeding, a momentary lapse of attention can disrupt the dialog. In the NISTARS case the worker would have to return to the computer display to retrieve the prompt or data. This kind of activity disrupts the concurrent activity organization of the system. More importantly it would tend to make the user uneasy about depending on voice and perhaps lead to using the computer display as a crutch for the voice system, degrading overall performance. Fortunately, supporting worker requests for voice output is a relatively easy feature to incorporate into a system which is already voice-interactive. Figure 3-5 shows part of a pick dialog (with machine output in a verbose form for readability) in which the user asks for a repetition of the pick quantity and then finds it necessary to determine what count had been entered before that entry was completed. Frequent requests to check on system output and own input would certainly degrade performance. However, without the ability to readily perform these functions by means of voice, the user's level of voice system acceptance could be compromised. The warehouse worker must be comfortable with voice interaction and feel that it can be depended upon even in the context of recognition errors.

Operator	Machine
	← “shelf three, box A two”
	← “enter stock number digits”
“two” →	
“four” →	
“enter” →	
	← “quantity ten of unit pair”
	← “enter count”
“repeat” →	
	← “quantity ten of unit pair”
“two” →	
“zero” →	
“echo” →	
	← “two zero”
“cancel” →	
“one” →	
“zero” →	
“enter” →	
	← “OK”

Figure 3-5 Requests for Voice Output

3.5.1 Error Handling

NISTARS work station computers already recognize the occurrence of input errors. Although most of these are handled as exceptions and have, therefore, been explicitly placed outside the voice-interactive dialog, errors associated with the voice hardware must be handled within the voice system. Reasonable handling of recognition errors is essential due to the "exacerbation process" (Delogu, Paoloni, & Pocci, 1991). When a user's utterance is not correctly recognized there is a tendency to alter the voice when entering the word a second time, leading to a second mis-recognition.

> When a name was not recognized (or not accepted by the recognizer in creating a template), many users would repeat the name with slow or exaggerated pronunciation as a habitual way of "being clearer". These habits hurt recognition; when the original template utterances had been spoken in a normal voice, exaggerated inputs were recognized less frequently than normal utterances. We observed two "levels" of this intrusive habit. Some users did not realize that exaggerated enunciation did not help. Others, who "knew better" found exaggeration difficult to avoid, both out of habit and as an automatic expression of annoyance when an utterance had been rejected. (Brennan, et al., 1991)

There is no way to completely avoid recognition errors and, even if they occur relatively infrequently, the cycle just described tends to make them highly frustrating to users. In order to increase users' confidence in the voice system, it is necessary to smoothly handle these errors when they do occur. In order to describe the approach to recognition error handling advocated here, it is necessary to sort out the range of entry errors which are possible. The list below includes all errors that can be expected to occur during the entry of a multi-digit number. The errors which are possible when, for example, a yes/no input is expected is a subset of the types presented.

1 failure to detect input

2 input detected but below recognition threshold

3 input incorrectly recognized, result is illegal type
(syntactically incorrect - "yes" when digit expected)

4 input incorrectly recognized, result is wrong instance of legal type
("one" for "nine")

5 input correctly recognized but result is input error on user's part
(typo - utterance was different from the one intended)

6 input correctly recognized but result is data error on user's part
(incorrect count)

Only type 5 and 6 errors exist in the current keyboard-based system. They are only detected when noted on the screen by the user or when the resulting input number is out of range as understood by the system. Error types 1 through 4 are failures or errors on the part of the recognition process. It is possible to imagine that type 1 errors occur currently when the user does not apply enough force during a key depression. However, such errors are usually detected by the user as the result of kinesthetic feedback. Under voice recognition, neither user nor system is immediately aware of type 1 errors. They are often detected when they produce syntax or data errors further along in the process. Some, such as a "yes" or "no" response, will leave the system and user both waiting for the next utterance in the dialog. It is therefore important to provide voice output in reaction to each voice input and when possible, have the system time-out on requests for simple responses. Although error types 3 and 4 are both termed "substitution errors" in the voice recognition literature, they must be handled in distinct ways. Type 4, like types 5 and 6, results in data entry errors which may or may not be caught by the system. If such errors occur at a high rate, it may be appropriate to echo input as part of the prompt which follows, to insure that incorrect data is not entering the system because it happens to be in range.

The system is capable of identifying both Type 2 and Type 3 errors. They are immediately identified as recognition errors and are potentially the most likely to initiate the exacerbation process. If the user was in the middle of entering a digit string when a Type 2 or 3 error occurred, the system should echo back those digits already entered in the same flat style preferred for input. It is hoped that speech modeling will help here to blunt the effect of the frustration cycle. If the Type 2 or 3 error occurred when the system was expecting a non-digit entry, the system should respond with an utterance such as "Expected yes, no" or "Expected enter"; again, the style should approximate that preferred for input. Errors of Types 4, 5 and 6 which result in an out of range input should produce a system response which echoes back the digits entered, such as "Rejected seven zero". The system should not read the example back as "Rejected seventy" as the user cannot enter the number in this fashion. If the error is, in fact, a recognition error, the digits should also be said in this way to take advantage of the modeling effect.

3.5.1 Voice Input Vocabulary

As presented in the sections above, the input or recognition vocabulary consists of seventeen words, as follows:

zero
one
two
three
four
five
six
seven
eight
nine

enter
yes
no
completed
cancel
repeat
echo

These words are proposed as sufficient to handle the task requirements of the carrousel work station functions. They are, however, only "proposed" since the confusability of this set of words cannot be determined without testing on the specific recognition system to be used. After the initial implementation of the system it may become necessary to change to alternate vocabulary items if any pairs result in high rates for substitution errors. This vocabulary size is well within the range of the hardware recognizers under consideration.

3.6 Template Training

Whether training an individual user's speaker dependent voice templates or building a set of speaker independent templates by training with a series of individuals, recognition requires training. (The cost trade-offs of using speaker independent versus speaker dependent recognizers for the NISTARS application are discussed in Section 5.) The recognizers under consideration all require word templates for comparison with the utterances heard during system operation. Training is generally a simple matter of prompting the user one or more times for each of the words in the vocabulary and testing the resulting templates by prompting again. The only *major design decision is the choice in prompting method* -- voice prompts or text prompts. It is certainly possible to provide minimally inflected voice prompts to stand as models for training input. However, training may be one situation where a reliance on modeling may not be appropriate. Providing the prompts in a text display may allow the users to concentrate on their own speech pattern (Mollarkarimi & Hamid, 1990). Since template training will be each user's introduction to speech recognition, it is also likely that not simultaneously introducing the novelty of voice output will smooth the entry into full voice interaction. When training with voice prompts, new users frequently begin their utterance before the system is ready to receive input. As a result, the input is often clipped at the beginning of the word, producing a bad template. Text prompts also avoid this problem and prevent the associated frustration from becoming many users' first impression of the technology.

4. VOICE INTERFACE IMPLEMENTATION PLAN

4.1 Integration Options

From the user interface point of view, the integration of voice with other technologies is relatively simple. The warehouse worker should always be able to switch from voice to keyboard entry, even in the middle of entering a series of digits. To the user therefore, the voice capability appears as a second and parallel I/O channel with reference to the keyboard/screen interface technology. The problem posed by the incorporation of voice hardware and software into existing NISTARS hardware and software systems is a good deal more complex. NISTARS installations vary in terms of hardware suite as well as installed software versions. DDD Jacksonville is the point of reference for the discussions which follow. Although system details vary between this and other sites, the main system outline for carrousel and hand-held operation is the same.

Control over carrousel functions, and therefore over the carrousel interface, is divided between the host Tandem computer and the PCs at the carrousel locations (referred to as workstation PCs below, to distinguish them from PCs playing other roles in the system). The Tandem plans worker trips, the sequence of orders to be picked or locations to be inventoried, and communicates each trip, step by step, to an individual PC. The workstation PC follows instructions from the host to display a specific screen configuration. Once the appropriate instruction and data have been passed to the PC, the PC takes control of interaction with the user. The PC processes keyboard inputs into field entries, performs data checking, and finally sends the completed transaction data to the host. Only the workstation PC knows the state of the interaction below the level of the current display form. The host sees results at the level of the filled-in form, except where exception processing may require additional communication. At DDD Jacksonville, existing workstation PC software is a mixture of Pascal and C, running under Xenix, although more recent versions of the software are written entirely in the C language.

In recent years, companies which produce voice recognition hardware and software have moved heavily into the area of telephone communications. Many new products have appeared for telemarketing and related applications. Although capabilities have been enhanced in the area of computer-based voice recognition systems, the hardware/software options have not expanded and in many cases have been reduced. Relevant off-the-shelf recognition products are limited to board level products for the IBM PC and clones, and are provided with drivers for DOS and C callable libraries. Several viable options exist for incorporating this technology into NISTARS systems, as described in the paragraphs which follow.

Figure 4-1 presents the simplest insertion of voice technology into the NISTARS system. Each workstation PC is equipped with a voice recognition board and attached headset microphone/speaker. This solution is the cheapest from the point of view of hardware costs since the hardware recognizers are the only additions to the system. The impact on software is not as attractive. Initially, DOS drivers must be ported to Xenix. This is a one-time cost and is relatively insignificant. The modification to the

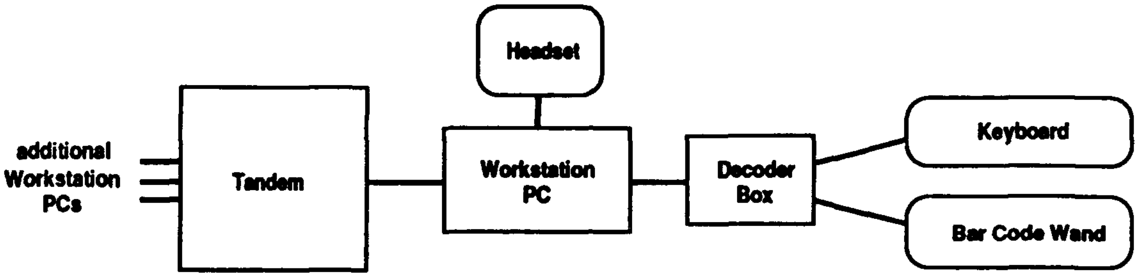

Figure 4-1 Single Station Voice Configuration

workstation PC software is more troublesome. Obviously the interface software must be modified to deal with both keyboard and voice recognition input. Unfortunately, the modified PC software must also deal directly with control of the voice recognition hardware. Calls and routines associated with a single vendor's recognizer product would become an integral part of the PC workstation configuration. Any change in the recognition hardware would necessitate revisions to the software on all workstation PCs in the system.

An open systems approach is appropriate in this context. Figure 4-2 shows voice technology integrated in the form of a "voice server". Server is not used here in the network client/server sense, but rather as a term to indicate a device which services a number of voice interface users and workstations. The voice server is a microcomputer which contains and manages a number of voice recognition boards,

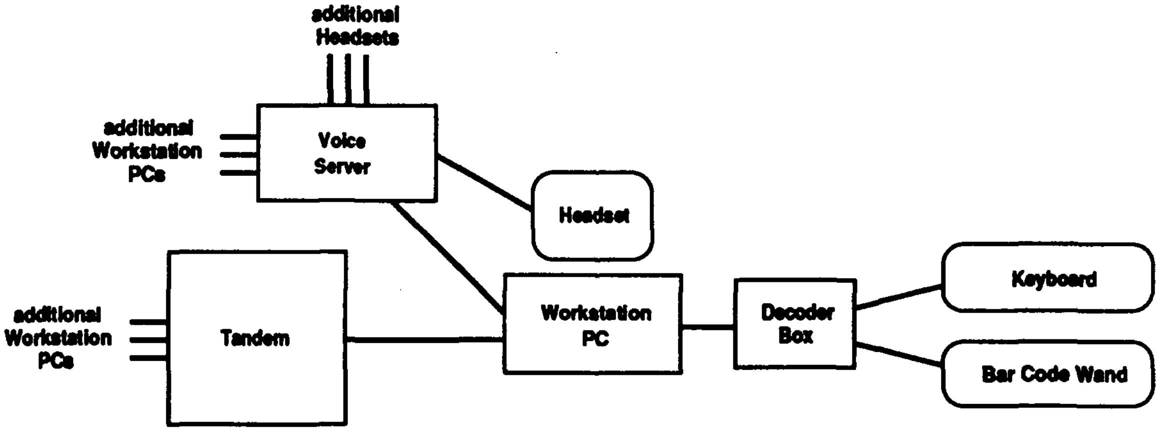

Figure 4-2 Carrousel Voice Server Configuration

each providing voice capabilities for a different workstation PC. The workstation PC communicates with the voice server in much the same way the Tandem host communicates with the workstation. Just as the Tandem sends a request for the workstation to display a form, the workstation may send the voice server a message to process an input field. Since the workstation PC communicates with the voice server by means of separate serial link, no additional load is placed on the Tandem/workstation communication line. Although the workstation's CPU will need to do more I/O processing, servicing an additional serial port, communication with the voice server will not disrupt communication with the host since interactions with the server will be within forms while interactions with the Tandem will generally occur between forms. Under this configuration, not only is the workstation divorced from the specifics of the voice hardware, but functions such as user requests for repetitions of voice output can be handled without the knowledge of the workstation. Dividing the voice interface processing between two machines provides a great deal of flexibility for voice system software design. Its prime advantage is, however, that modifications can be made to the voice interface hardware and software without modifying the workstations themselves. Changing between speaker-dependent and speaker-independent hardware or between competing vendors from installation to installation would require only changes to the voice server software.

The incorporation of voice technology into the fixed carrousel workstations is not intended as the final application of voice technology in NISTARS. Although the carrousel workstations were selected here as the best opportunity for immediate implementation, extending the interface described above to the hand-held stations in the near future is certainly a realizable goal. The voice server strategy, and even the voice server software, can be extended to this environment with little change. Figure 4-3 presents an overview of a configuration for handling both carrousel and hand-held stations. Since the only external access to the hand-held is via the RF channel, the server is shown connected through the RF base station. The NISTARS RF configuration is very complex, but supports maximum flexibility for equipment usage. Through a series of base stations, utilizing polling strategies to multiplex multiple hand-helds on a single RF channel and search through multiple channels, any hand-held unit can be used at any location for any function. The complexity of this arrangement would require a significant design effort to support the RF connection of headsets and servers in a similar manner. In order for voice to be generally usable in the hand-held system, headset units must be able to float within the warehouse facility in the same way as the hand-held units themselves. Not only does the concept of a voice server transfer directly to this configuration but the available independent computational power will allow for the control necessary to associate any headset with any hand-held. Implementation of the voice server in the carrousel context will not only introduce the technology in a simpler hardware/software environment but also, with user acceptance and performance enhancement established, pave the way for the more extensive effort required to integrate voice in the hand-held stations.

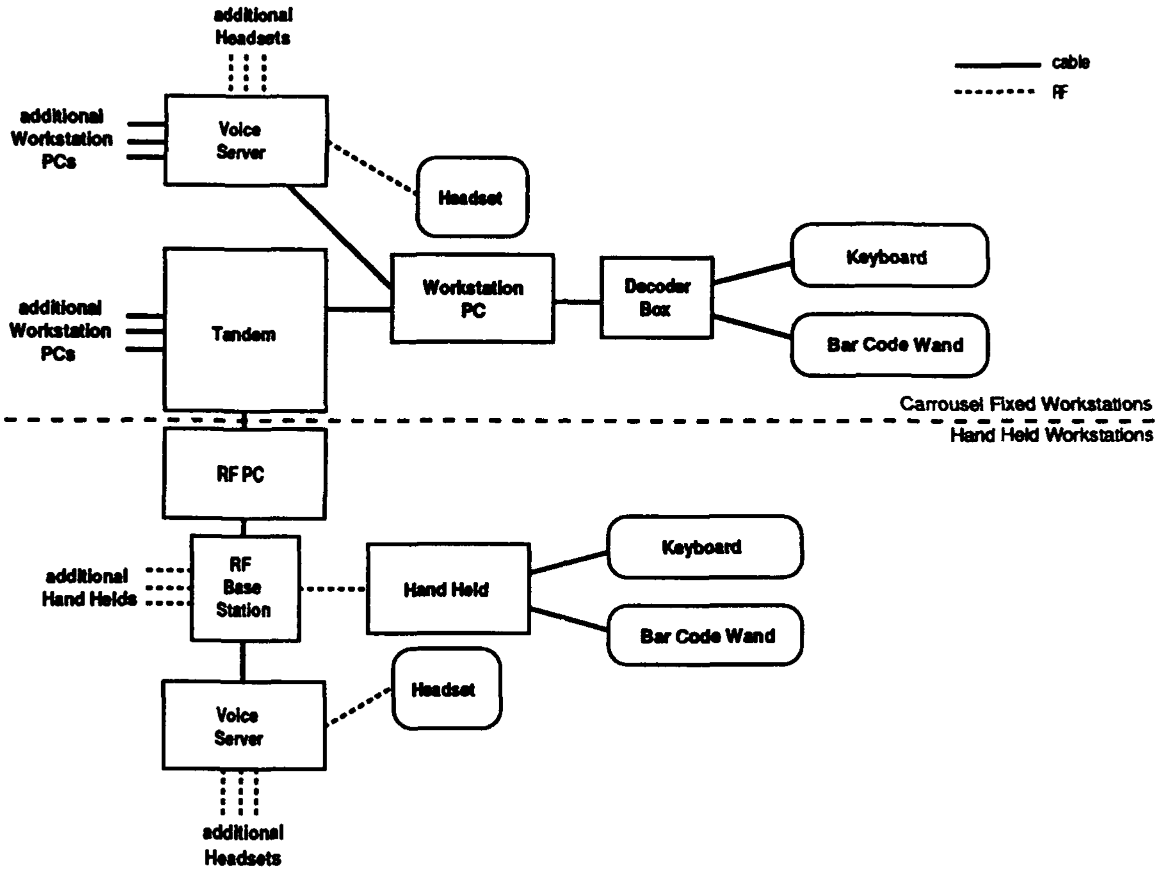

Figure 4-3 Carrousel and Hand Held Voice Configuration

4.2 Voice Server Specifications

The delivery platform for the fixed carrousel station voice server will be an IBM PC or clone equipped as follows:

- 80486DX processor
- 4 MB RAM
- 20 MB hard disk
- minimum of 4 voice recognition boards
- minimum of 4 serial ports
- VGA card
- MS-DOS 5.0

Dedicated monitors and keyboards should not be required for the servers. The video card is included so that maintenance personnel can perform routine update tasks and run diagnostics and performance tests without opening the machine. The existing workstation PCs have available serial ports so no modification is necessary for compatibility with the serial ports planned for the server.

Several items in the list are contingent on characteristics of the voice hardware and completed interface which cannot be determined at this time. For example, it is intended that all executable server code be resident in memory as well as the digitized form of all voice output messages. Timing constraints require that digitized messages be directly available in memory for transmission to the voice boards. Depending on the algorithms utilized by the vendors development software, the memory required to hold the digitized messages could conceivably allow a reduction or force an increase in required RAM. Similarly, it is possible that the requirement for four serial ports could be satisfied by the addition of a two port card and utilization of onboard ports. It is more likely, however, that time constraints will require the incorporation of a multi-port smart serial card which can perform such tasks as buffering without tying up the CPU and degrading voice response time. The hardware requirements noted here and utilized in the section on cost below are conservative and may well be reduced in practice.

4.3 Task Breakdown

Under an SBIR Phase II effort all necessary software for workstation and server PCs can be developed and a complete multi-server system installed at a single NISTARS site. In order to bring the first voice server carrousel system on line, CHI Systems will follow the development plan specified by the following tasks :

Task 1: Design Workstation/Server Software Architecture

The central design activity under Task 1 is to establish the division of labor between workstation and server. Major questions include whether or not all syntax checking will be done on the server. For example, will the workstation request specific types of voice entry such as number-input, "yes"/"no"-input or a "completed"-input from the server, or will it simply accept whatever recognition results are passed to it? If the syntax checking strategy is pursued, will the server also perform range checking in order to facilitate rapid voice response? If an extended hand-off of input field processing of this sort is to be developed, the interleaving voice and key entry becomes a potential problem. If a number entry was begun via voice but completed by key, care must be taken to assure that the server does not get out of synch with the workstation state. These and other problems must be resolved and the solutions formalized in the design for workstation/server communication - the protocols for message passing between the two devices.

Task 2: Design, Implement, and Test Workstation Modifications

Since the various NISTARS sites are operating under different versions of the workstation PC software it is essential that a single site, for prototype testing and installation, be established early on in the effort. Once a site is selected under this task, the workstation software from that site must be obtained from the Navy. (Although DDD Jacksonville has been the focus of the current Phase I effort, it is unlikely that the Jacksonville facility can be used as the prototype site since it is not running the latest version of the relevant software.) Analysis of the existing software will proceed in parallel with the design effort under Task 1. Equipment comparable to the existing workstation hardware will be acquired, as well as the necessary software environment and tools to modify the fixed carrousel station software. Modifications will be designed, implemented, and tested at CHI Systems, incorporating mockups of the Tandem communication as necessary.

Task 3: Design, Implement, and Test Server Software

The first step in the development of the server software will be the acquisition of the hardware and software necessary to build a single server, complete with multiple voice recognition boards. In addition to the usual development tools such as compilers, it will also be necessary to acquire those tools provided by the voice hardware vendor, such as digitization software. Design of the server software will begin after the Task 1 communications design effort is well underway. A major focus of the Task 3 design effort will be to establish servicing priorities for the various voice inputs, voice outputs, and workstation communications. For example, if a recognition has occurred and a message from a workstation is pending, which should be handled first? Or further, if the message and pending recognition both belong to the same workstation, what action should be taken? Once established, software must be designed and implemented to support the servicing priority scheme and tested with a full complement of recognizers. The prototype server software developed under Task 3 must include data collection software for assessment of server performance.

Task 4: Install Prototype Voice-Based Station at NISTARS Site

Under Task 4, the fully implemented prototype with data collection capability will be installed on site at the selected NISTARS facility. The modified workstation software will be loaded onto a single carrousel workstation PC. CHI Systems will work closely with warehouse and contract personnel to verify hardware/software functionality and compatibility and develop a test plan to exercise the equipment. Selected workers on the warehouse floor will be trained on the system and allowed to perform voice-based carrousel functions as specified by the completed test plan.

Task 5: Collect System Performance and Acceptance Data

Information will be gathered on all aspects of prototype functionality. Use of the system by warehouse personnel will be observed in order to evaluate the effectiveness of training and the habitability of the voice dialog as designed. The observation of problems as they occur is of particular importance for the development of solutions. The prototype version will also collect performance data in order to assess interface effectiveness in a non-subjective manner. Upon completion of test use of the voice system, workers will be interviewed to gauge the level of user acceptance and system responsiveness and to identify problem areas for further analysis.

Task 6: Evaluate Results and Modify System As Required

Upon completion of Task 5, all data resulting from the prototype testing will be analyzed. Server and workstation software will be modified as required by the knowledge gained on site. Based on the final configuration, hardware for multiple servers will be acquired. It is possible that information acquired during development and testing will allow relaxation of the server hardware/software requirements and lead to the acquisition of cheaper platform configurations for the final operational systems. The final workstation/server configuration will be fully tested at CHI Systems before delivery to the Navy.

Task 7: Install Operational Voice-Based System at Site

User-level and software maintenance-level documentation will be prepared for the final version of the system. Servers and workstation software will be installed at the selected facility, providing voice interfacing for all fixed carrousel work stations on site. CHI systems will train all relevant computer, management, and floor personnel and support the installation as required by the facility.

5. COST BENEFIT ASSESSMENT

5.1 Overview

The proposal for the current effort stated that the equivalent of four separate system designs would be examined for costs and benefits:

1. speaker dependent with visual feedback,
2. speaker independent with visual feedback,
3. speaker dependent with auditory feedback, and
4. speaker independent with auditory feedback.

As described in Section 3.1 above, for voice to be an effective interface technology in the context of NISTARS procedures, the visual feedback option is not a viable alternative. Therefore versions 1 and 2 in the above list are known to provide no benefit and need not be discussed further. At the same time, versions 3 and 4 above have not resulted in the production of distinct designs as originally anticipated. This is in large measure due to the characteristics of the currently available recognizers considered fit for the application under consideration.

The two principal candidates for use in the voice server system are a speaker independent recognizer made by Scott Instruments Corporation and a speaker dependent recognizer made by Votan. As expected, the speaker dependent system requires the development of user training software as well as requiring each user to dedicate time to establish their own set of recognition templates. Unlike earlier speaker independent systems, Scott Instruments provides the software for developers/users to create their own speaker independent templates. Their algorithms are apparently robust enough to dispense with the massive efforts previously required to create speaker independent vocabularies. Scott Instruments is not in the vocabulary template development business although speaker independent templates for the digits and "yes" and "no" are provided with the development package. Instead, the developer is expected to run a sizable sample (75 individuals) of potential users through a training process like that used for training speaker dependent systems. Once the templates are established, new hires need not go through the training pass, but the templates can be updated if one or more individuals find their recognition results to be under par. As a result of Scott Instruments' approach, the speaker dependent and speaker independent alternatives for this project have converged. Both place requirements for training on the developer as well as the end user. Although users will be required to train at different times, the end result will probably be very similar in terms of overall costs.

It is frequently observed that "almost every desirable capability (e.g., speaker independence, continuous speech, and rejection) also degrades the accuracy of a system." (Lee, et al.1990) As a result of these tradeoffs, the two principal candidates for use in the voice server system differ in their ability to handle continuous or connected speech. The speaker independent system requires the user to clearly separate words when speaking, the speaker dependent system does not. Although this was not considered a major distinction in the proposal, the fact that other distinctions have been considerably reduced and the fact that the interface will be highly dependent on users' ability to quickly enter digit sequences, it is a point to be considered. Unfortunately, only a comparison of both candidate recognizers performing in the carrousel workstation prototype could determine if continuous recognition confers a significant advantage in this case.

The interface design and server implementation plan presented above provides the best means available to apply voice technology, whether speaker dependent or speaker independent, to the unique requirements of the NISTARS warehouse. As a result, the benefits from using either of the two systems will be identical, with the possible exception just noted. At the same time, the voice interface design presented for the carrousel work station does not represent a drastic change from existing procedures. Perhaps of more interest in the sections below is not which recognizer type is more cost effective, but rather if any automated voice system is cost effective when compared to the existing manual/visual system.

5.2 Expected Benefits

The steps taken to locate and retrieve a container from the carrousel and to verify location and container contents to the system are the same for each of the carrousel functions. In order to assess the gains made by voice over the existing system, the voice and manual/visual versions of this front-end procedure were compared. Each procedure was acted out repeatedly and timed. Since actual utterance and response times can only be guessed, the acting out was based on previous experience with similar devices. Results are probably accurate within a one or two second range. The performance of the existing manual/visual procedure consistently took 16 seconds. The voice-based procedure varied from 7 to 8 seconds in duration. Since there is likely to be some small additional saving later in each procedure, the overall savings may be estimated at 9 seconds per task. The raw size of the time saving is of course of little value and is only useful when viewed relative to total task time.

During the course of the effort reported here, a copy of the August work log printout was obtained from DDD Jacksonville. The appendix of this report presents the log-on time, number of tasks (both taken directly from the log) and average number of minutes per task calculated for each worker as reported for each function type. Table 5-1 presents a summary of that information. Unfortunately, carrousel functions are under-represented in the sample (most of the binnable aisles are reported as non-mech as well). This is an artifact of the way in which the work log report is compiled. The log only includes cumulative statistics for those individuals who worked on the day the report was generated. The numbers are presented here since they are the only hard information available. As the time per task figures have been calculated using log-on times, the task duration shown is particularly undependable at the high end where the numbers may reflect nothing more than a worker's failure to immediately log off after completion of the trip. Since the figures cannot err in the direction of estimating too little time per task, the low end figures are probably closer to the truth. In response to a query from CHI Systems, one warehouse manager offered 55 picks an hour (1.1 minutes per task) as a reasonable rate for work on the carrousels. One example from

Table 5-1 Work Log Summary

Station/Task	Total Hours	Total Tasks	Minutes Per Task
Carrousel Picks	19.96	796	1.5
Binnable Picks	18.63	585	1.9
Non-Mech Picks	636.56	16,197	2.4
Carrousel Inventory	30.22	777	2.3
Binnable Inventory	11.07	378	1.8
Non-Mech Inventory	100.82	3,558	1.7

the log, 43 issues in 0.83 hours, comes very close to that figure at 1.2 minutes per task or 50 picks per hour.

The inventory function seems to take somewhat more time, as seen in the log figures (stow is not broken out for carrousel in the work log) and the corresponding time saving will be smaller relative to task duration. Since carrousel tasks thus appear to vary in a range from one to one and a half minutes in duration, a conservative estimate of the time saving to be expected from the application of voice is on the order of ten percent or an eleven percent increase in production. Again, the raw figure for percent improvement is valuable information but without data on transaction volume the true benefit cannot be determined. Unfortunately volume figures are not available in any direct fashion since there are apparently no standard reports generated which break out the tasks of interest. Contract computer personnel at DDD Jacksonville report that the facility processes from 1000 to 1200 issues per day, 31 to 32 percent of which are carrousel issues (based on examining figures covering a four month period). Since the number of stows and inventory tasks will be considerably lower, we may conservatively estimate 350 carrousel transactions per day. This converts to only six or seven hours per day, spread across ten fixed carrousel stations, a fairly low level of usage (less than 10% utilization per station based on an eight hour work day).

Although voice may enhance system accuracy by removing the need for workers to look away from the carrousel at several points while retrieving containers and by providing a verbal readout of.the full NSN for comparison during inventory, the principal benefit of implementing a carrousel voice interface is the resulting reduction in per task time. At critical points during various military operations the reduction in time may prove to be very valuable. However, dollar savings over time can only be calculated based on transaction volume. If projections for carrousel work station use system-wide are significantly higher than the volume which can be discerned at DDD Jacksonville, a ten percent increase in production would prove cost effective, as described in the section which follows.

5.3 Expected Costs

The speaker dependent recognizer market ranges from devices available for under two hundred dollars to those costing ten thousand and more. The low end of this range basically qualifies as hobbyist equipment while the high end consists of large vocabulary recognizers which are of little use in the NISTARS application. Votan's Standard Voice Card, priced at $1,500 in quantities of two or more, is the candidate speaker dependent device selected for cost comparison. Although a cheaper TI board provides similar functionality, past experience indicates that the Votan board provides better recognition in the context of industrial noise and is generally a superior product. Higher priced speaker dependent recognizers begin at nearly double Votan's price and are already the low end of the large vocabulary (1000+ word) systems.

The speaker independent market presents a narrower range, running from fifteen hundred dollars to five thousand. The Scott Instruments Corporation Model 20 Recognition Processor, priced at $1,495, is the candidate speaker independent device selected for cost comparison. Although it represents the low end of the of the speaker independent market, its functionality is comparable to the Votan card with the exception of speaker independence. Votan and Scott Instruments have both been in the voice recognition business since the early 80's and have continued to develop and improve their products, each of which can be described as state-of-the-art.

In order to allow the warehouse worker freedom of movement for a task that involves a great deal of physical repositioning, it is recommended that the selected recognizer be provided with a locally RF linked microphone/earphone headset. Both microphones and RF systems vary widely in capability and price. Without on-site testing it is impossible to determine how a particular product will perform in the noise and RF environment of a NISTARS warehouse. Discussions with recognizer vendor technical personnel indicate that RF headsets in the three hundred and fifty dollar range should be adequate for the job.

The 486 computer to be used as the server delivery vehicle, described in Section 4.2 above, can be acquired for under fifteen hundred dollars. A price of $1,425 was developed from published sources. The addition of a multiport smart serial card will cost eight hundred dollars. Based on these figures, a fully equipped, four recognizer server costs out to $9,625, with a cost per work station of $2406.25. Since all items are likely to be available at a 10% to 20% reduction in price if purchased in large quantities, this figure can probably be lowered in practice.

Assuming for the moment a rounded up cost of $2500 to provide voice capability to a carrousel work station and also assuming an hourly labor cost of ten dollars, the upgrade to voice would have to save 250 hours of labor before paying for itself. With a predicted 10% time saving with voice interfacing, a pre-voice workstation utilization time of 2500 hours would be required. Over a two year period this represents approximately five hours per work day which is over 50% utilization of the workstation, well above the numbers currently available from DDD Jacksonville. In general, the break-even point can be calculated by utilizing the estimate of a nine second saving per transaction described in Section 5.2, and rounding the year down to 250 working days. In order to save the 250 hours of labor required to break even, the following transaction rates per workstation are required:

1 year	=	400 transactions per day
2 year	=	200 transactions per day
3 year	=	134 transactions per day
4 year	=	100 transactions per day

Although no hard information is available as to the durability of the voice equipment, it is of interest to note that Scott Instruments publishes an MTBF of 100,000 hours (calculated) power-on time for its Model 20 speaker independent recognizer. It should therefore be expected that even at lower transaction rates the voice hardware will still be on line when the break-even point is reached.

Since all estimates to this point have been intentionally conservative, it is fair to assume that reductions in cost based on quantity purchase and modified requirements for the operational system could bring even Jacksonville to the level of a three year saving of replacement cost. In the near term it can also be expected that hardware costs will come down while labor costs will go up. Since the maintenance on recognition hardware is minimal and the PCs can be covered under existing maintenance policies for minimal cost, at the level of equipment acquisition and maintenance a voice system can be marginally cost effective immediately. The magnitude of this cost effectiveness is also expected to increase over time if the volume of processing is increased or the time required for the user to engage in operations other than computer I/O (e.g., material handling) is decreased.

It is assumed that all system and software development costs will be covered under a Phase II SBIR. The justification for this effort must be premised on the long term impact of incorporating voice in NISTARS. Although it is impossible at this time to gauge the cost of design and implementation of an RF system which can support multiple voice servers, it is certainly the case that the server as developed for the fixed carrousel stations can be used virtually unchanged for a much larger system incorporating all hand-held stations. It is also true that one of the main reasons for the high cost of incorporating voice is the lack of consideration of this technology in NISTARS evolution to date. If voice interfacing for carrousel stations is taken as a first step to phase in voice technology, future developments of NISTARS voice applications will be more profitable. High payoff areas such as cold storage hand-helds will require minimal additional development for much greater time gains. Higher utilization rates for existing carrousel stations is also a near term possibility as various economic pressures may result in facility consolidations.

A single cost area requires further comment - template storage. A speaker independent system with multiple voice servers can be supported by storing the master template set on each server. A speaker dependent system with multiple voice servers requires centralized storage so that an individual's templates can be downloaded or trained and uploaded from any server in the system. The Tandem is the only machine in the warehouse system which can manage this kind of centralized data repository. Unfortunately the Tandem is the central processing unit in a highly complex hardware and software configuration. Since this configuration is already developed and maintained by contract computer/software personnel, it is not appropriate for CHI Systems to design or propose to design and implement changes that go to the heart of the system. The Navy would be better served by having the existing NISTARS contractor perform this work if it proves necessary. It is sufficient here to note that the task is non-trivial and would probably require additional communications hardware and software at the warehouse and server levels.

5.4 Recommendations

Unfortunately there is no available independent assessment of the relative recognition accuracy of the Votan and Scott devices. Nor is there any way to predict the impact of continuous recognition of digits as compared to discrete recognition. Both areas could be assessed relatively cheaply as part of a Phase II effort, but it is likely that the cost of providing centralized storage for speaker dependent templates would overshadow the differences in these areas. As the two alternatives are otherwise very similar from a cost/benefit point of view, it is CHI Systems' recommendation that the voice server be implemented with Scott Instruments recognition hardware. Based on the information available during the course of this effort, it has been determined that incorporating voice technology into the fixed carrousel work stations is marginally cost effective at present. However, it must also be noted that the relative benefit of the voice interactive feature is sensitive to both the time required for functions other than computer I/O and to the volume of processing performed with the system. Substantial changes in either matenal handling procedures or in the total volume of transactions could significantly increase the cost-effectiveness of voice. It is recommended that the voice server approach be implemented now in order to establish voice interfacing as a viable alternative within NISTARS. Continued development of the NISTARS system without attending to the potential inherent in voice technology will make it more difficult to incorporate this technology to advantage in the future.

REFERENCES

Brennan, P., Deffner, G., Lawrence, D., Marics, M., Schwab, E., and Franzke, M. Should we or shouldn't we use spoken commands in voice interfaces? Reaching Through Technology: CHI '91 Conference Proceedings, pp. 369-372, April-May 1991.

Delogu, C., Paoloni, A., and Pocci, P. New Directions in the Evaluation of Voice Input/Output Systems. IEEE Journal on Selected Areas in Communications, vol 9, no. 4, pp. 566-573, May 1991.

Jones, D., Hapeshi, K., and Frankish, C. Design guidelines for speech recognition interfaces. Applied Ergonomics, vol. 20, no. 1, pp. 47-52, March 1989.

Karis, D., and Dobroth, K. Automating Services with Speech Recognition over the Public Switched Telephone Network: Human Factors Considerations. IEEE Journal on Selected Areas in Communications, vol. 9, no. 4, pp. 574-585, May 1991.

Lee, K., Hauptmann, A., and Rudnicky, A. The Spoken Word. Byte, vol. 15, no. 7, pp. 225-232, July 1990.

Leiser, R. Improving natural language and speech interfaces by the use of metalinguistic phenomena. Applied Ergonomics, vol. 20, no. 3, pp. 168-173, September 1989.

Mollarkarimi, C., and Hamid, T. Remote Voice Training: A Case Study on Space Shuttle Applications, Appendix C. Lockheed Space Operations Co., Cocoa Beach, FL. Report No.: NASA-CR-187385, 1990.

Zoltan-Ford, E. Reducing Variability in Natural-Language Interactions with Computers. Proceedings of the Human Factors Society 28th Annual Meeting, pp. 768-772, October 1984.

Appendix

DDD Jacksonville Employee Statistics

Carrousel Picks

Total Hours	Total Tasks	Minutes Per Task
0.04	3	0.8
0.83	43	1.2
0.11	5	1.3
11.92	526	1.4
0.37	13	1.7
1.08	38	1.7
2.59	87	1.8
2.19	61	2.2
0.83	20	2.5

Carrousel Inventory

Total Hours	Total Tasks	Minutes Per Task
0.20	8	1.5
6.16	185	2.0
2.72	79	2.1
10.40	285	2.2
1.06	29	2.2
1.70	46	2.2
1.05	28	2.3
6.93	117	3.6

Binnable Picks

Total Hours	Total Tasks	Minutes Per Task
0.03	1	1.8
18.24	579	1.9
0.36	5	4.3

Binnable Inventory

Total Hours	Total Tasks	Minutes Per Task
11.00	376	1.8
0.07	2	2.1

Non-Mech Picks

Total Hours	Total Tasks	Minutes Per Task
1.09	85	0.8
0.18	14	0.8
0.27	20	0.8
1.61	91	1.1
11.05	618	1.1
4.05	190	1.3
7.68	341	1.4
0.67	28	1.4
4.13	165	1.5
2.64	97	1.6
0.25	9	1.7
36.72	1282	1.7
29.71	1021	1.7
9.77	298	2.0
1.72	51	2.0
2.67	79	2.0
40.42	1146	2.1
54.85	1541	2.1
30.35	845	2.2
42.25	1121	2.3
21.86	572	2.3
0.08	2	2.4
64.38	1600	2.4
21.84	531	2.5
21.78	510	2.6
0.52	12	2.6
0.44	10	2.6
8.67	183	2.8
70.89	1465	2.9
32.04	654	2.9
15.77	315	3.0

Non-Mech Picks (continued)

Total Hours	Total Tasks	Minutes Per Task
9.26	183	3.0
2.06	38	3.3
5.35	93	3.5
8.00	111	4.3
0.44	6	4.4
48.52	645	4.5
3.04	38	4.8
0.68	8	5.1
11.46	134	5.1
1.63	16	6.1
2.41	23	6.3
0.13	1	7.8
3.23	5	38.8

Non-Mech Inventory

Total Hours	Total Tasks	Minutes Per Task
0.02	3	0.4
0.73	69	0.6
1.20	72	1.0
0.53	31	1.0
2.37	136	1.0
4.11	232	1.1
3.81	215	1.1
0.15	7	1.3
1.75	78	1.3
7.79	338	1.4
2.52	108	1.4
2.69	106	1.5
1.07	42	1.5
16.55	611	1.6
0.25	9	1.7
1.34	47	1.7
2.58	89	1.7
5.90	201	1.8
0.03	1	1.8
0.55	17	1.9

Non-Mech Inventory (continued)

Total Hours	Total Tasks	Minutes Per Task
21.59	660	2.0
0.10	3	2.0
1.88	53	2.1
3.66	92	2.4
1.59	35	2.7
13.17	273	2.9
0.77	15	3.1
0.16	3	3.2
0.51	6	5.1
1.45	6	14.5