THE HUMAN FACTORS OF AUTOMATIC SPEECH RECOGNITION IN CONTROL ROOM SYSTEMS

CHRISTOPHER BABER

Doctor of Philosophy

The University of Aston in Birmingham

August 1990

For my Parents

Human Factors of Automatic Speech Recognition in Control Room Systems

Christopher Baber

Doctor of Philosophy 1990

This thesis addresses the viability of automatic speech recognition for control room systems; with careful system design, automatic speech recognition (ASR) devices can be useful means for human computer interaction in specific types of task. These tasks can be defined as complex verbal activities, such as command and control, and can be paired with spatial tasks, such as monitoring, without detriment. It is suggested that ASR use be confined to routine plant operation, as opposed the critical incidents, due to possible problems of stress on the operators' speech.

It is proposed that using ASR will require operators to adapt a commonly used skill to cater for a novel use of speech. Before using the ASR device, new operators will require some form of training. It is shown that a demonstration by an experienced user of the device can lead to superior performance than instructions. Thus, a relatively cheap and very efficient form of operator training can be supplied by demonstration by experienced ASR operators.

From a series of studies into speech based interaction with computers, it is concluded that the interaction be designed to capitalise upon the tendency of operators to use short, succinct, task specific styles of speech.

From studies comparing different types of feedback, it is concluded that operators be given screen based feedback, rather than auditory feedback, for control room operation. Feedback will take two forms: the use of the ASR device will require recognition feedback, which will be best supplied using text; the performance of a process control task will require task feedback integrated into the mimic display. This latter feedback can be either textual or symbolic, but it is suggested that symbolic feedback will be more beneficial.

Related to both interaction style and feedback is the issue of handling recognition errors. These should be corrected by simple command repetition practices, rather than use error handling dialogues. This method of error correction is held to be non intrusive to primary command and control operations. This thesis also addresses some of the problems of user error in ASR use, and provides a number of recommendations for its reduction.

KEYWORDS

Human Factors; Automatic Speech Recognition; Process Control; Feedback; Training

Acknowledgements

I am indebted to my supervisors, Dr. R.B. Stammers at Aston University, and Dr. D.M. Usher previously with the CEGB, for their assistance, comments, criticism and support. Although many of the ideas presented in this thesis are the result of fruitful collaboration between us, Dave and Rob cannot be held responsible for them. Any errors, mistakes, and misconceptions are the fault of the author. Many thanks too, to my colleague Neville Stanton, whose comments and criticism have helped shaped this thesis.

Many of the studies in the thesis have relied on software written for me, and for which I am very grateful. I would like to thank Ray Taylor of Aston University, Dave Usher (again), and Mike Carey formerly of Aston University for their patience, fortitude and, above all, programming talents.

The data for studies one, four, six and nine were collected in collaboration with Final Year students at Aston University, and I would like to express my gratitude to: Sally Aitken, Bina Bharadia, Ashley Rosengarten, and Ady Yong.

The staff at Durley Park Control Centre were extremely helpful in their participation in the assessment of the telecommand demonstration, as were the staff and students at Matthew Boulton College, Birmingham, who assisted in the feedback study and the staff at British Gas,West Midlands Research Station, who participated in the workload study. In particular, I wish to thank Brian Johnson, of Durley Park, Jim Turberville of Matthew Boulton College, and Gerry Brennen of British Gas for their assistance in bullying, badgering and cajoling 'subjects' for the studies.

Much of the work reported in the thesis involved travel to Bristol. I am very grateful to Dave, Sue and Alice Usher for their hospitality during these visits.

TABLE OF CONTENTS

Page

CHAPTER ONE

CHAPTER TWO

CHAPTER FOUR (continued) Page

CHAPTER FIVE

CHAPTER SIX

CHAPTER SIX (continued) **Page**

CHAPTER SEVEN

CHAPTER ELEVEN

CHAPTER TWELVE

CHAPTER TWELVE (cont.) Page

Page

CHAPTER THIRTEEN

CHAPTER THIRTEEN (cont.) Page

CHAPTER FOURTEEN

CHAPTER FOURTEEN (cont.) Page

CHAPTER FIFTEEN

TABLE OF FIGURES

TABLE OF FIGURES (cont.)

TABLE OF FIGURES (cont.)

CHAPTER ONE

INTRODUCTION

The background to the research contained in this thesis is briefly described. Automatic Speech Recognition is presented as a novel, but misunderstood technology. Some of the myths surrounding it are discussed and its real potential presented. Further, the reasons for concentrating on Automatic Speech Recognition at the expense of other examples of speech technology, such as speech synthesis, are presented. The Introduction ends with brief precis of each of the eleven main chapters in the thesis.

The Research Project

This thesis is the result of a three year project, conducted in collaboration with the CEGB on a SERC Case Award Scheme. The initial goal of the project was to develop a set of guidelines and potential areas of application for speech technology in all aspects of CEGB operation. However, during the course of the project this goal was revised for a number of reasons.

The area of speech technology is very broad (as the definition section below illustrates). It did not seem feasible to address the human factors aspects of all types of speech technology in three years. Therefore, it was decided that automatic speech recognition (ASR) receive most attention. This decision was arrived at from the consideration of input devices in an earlier collaboration between Aston University and the CEGB, and from the realisation that other forms of speech technology would not be practicable for control room operations.

A further factor which resulted in a change of research emphasis was the privatisation of the CEGB, by the British Government during 1988-1990. This led to the division of the CEGB into four separate companies. Of greatest import to this thesis was the subsequent confusion of company identities, and serious problems of access to sites.

Therefore, the study of applications in the field had to be greatly reduced from those intended. However, this in turn led to more emphasis being placed on the laboratory based research techniques of human factors and cognitive psychology. Consequently, the general goals of this thesis are four fold:

i.) To assess the potential of ASR in Control Room Systems;

ii.) To define possible areas of application, in terms of Control Room Tasks and the limitations of Human Operators;

iii.) To provide empirically derived guidelines to support such application decisions;

iv.) To define the nature of speech based interaction with computers in Control Room Systems.

Speech Technology: Myths and Potential

Computers which can converse intelligently with humans have long been the staple diet of science fiction. There have been a number of films and books which present computers and robots that can either converse with people (such as "Hal" in 2001 or "Marvin the Paranoid Android" in Hitchhikers Guide to the Galaxy) or which can understand speech (such as "R2D2" in Star Wars).

Such a presentation tends to polarise the views of the audience; they either assume that the problems of speech technology have long been solved, or that speech technology is something which exists only in the realms of science fiction. The well publicised mishaps with speech technology, such people 'disabling' loudspeakers in 'talking lifts', only serve to strengthen this latter point of view. Thus, whichever view is held by prospective users, it would be reasonable to expect a degree of skepticism, as to the viability of speech technology.

When one thinks of speech as a means of communication, one will tend to think of it in conversational terms, which, after all, is its most common use. This use of speech reflects a highly developed human skill which computers are unable to mimic (Newell, 1984). Many promoters of ASR celebrate the 'naturalness' of human speech, and claim that speech technology can exploit this naturalness. However, as Teja and Gonnella (1983) point out,

> "Human speech is a communication medium that is natural for people, but not one that is best suited for machines."

A computer which could hold a conversation with its user would indeed be revolutionary, but would it offer desirable benefits? Do people want to chat with their computer? Primarily, people hope to be able to speak directly to the computer and for it to obey their commands, not answer back. This raises the interesting point that speech can be used in a variety of communication styles (Potter and Wetherell, 1987). Exactly how people will want to 'converse' with computers will depend on the nature of the task and the environment in which the task is performed.

Some writers (e.g. Helander et al, 1988), suggest that ASR is natural, not simply because it uses speech, but because it does not require the learning of a motor skill, such as typing. The dream of being able to use a computer without any previous training is tempting, but people are very rarely able to use any form of speech technology without some degree of training and practice. So much for the claims that using speech will be 'natural'. One might now ask, what is speech technology and why should it be considered at all?

To answer the second question first, most activities in the Control Room rely on either manual action, such as typing, or the acquisition of visually presented information from a VDU or mimic diagram. This has the potential to 'overload' the operators' visual information channels. The introduction of speech technology could allow these channels to be unburdened by employing verbal information. Furthermore, the use of speech technology could allow tasks and activities to be restructured to produce easier and more efficient patterns of work.

Definitions of Speech Technology

The U.K. Department of Trade and Industry has recently published a document entitled, "Speech and Language Technology - Strategy for Research and Development Support" (March, 1989). In this document, speech technology is divided into three main sections: speech input, speech output, and speech coding. Of these, the latter is of minimal interest to the concerns of this thesis. Therefore, speech technology will be defined as either speech input to or output from computers.

i.) Speech Input

Speech input refers to the processes by which human speech is received and processed by a computer. This suggests a two stage process of reception and processing.

An initial distinction can be drawn according to the type of processing the computer performs on the received speech. The computer could be using the speech to match against previously recorded speech of the current user, in order to identify the user. Such **speaker identification**, or **speaker verification**, is available in several commercial products. Indeed, in some States in America, speaker identification equipment can be used to provide evidence in Courts. This process is assumed to be similar to finger printing, in that each person has a distinctive 'voice print'. However, the concept of speaker identification is not without its critics (see Nolan, 1983 for a comprehensive review). Security is an important consideration for many aspects of control room operation. However, it is proposed that the techniques of speaker identification will not offer any additional benefits to existing security practices.

A second form of processing will require the computer to perform specified responses to human speech. In this form of speech input, the computer is said to recognise the speech. **Speech recognition** can vary in complexity along several dimensions, according to the way samples of speech are collected, the number of speakers the device can accomodate, the number of words in its vocabulary, and the complexity of the processing, e.g. whether the device simply responds to speech or whether it must perform further processing to 'understand' the speech.

Speech recognition is currently capable of recognising speech with sufficient accuracy for it to be used in several industrial applications (see chapter six). Commercial devices tend to rely on variations on the theme of filter bank analysis, but over the past decade statistical techniques have been developed by researchers to provide better performance. The most common form of commercially available ASR devices require users to provide samples of their speech to create 'templates', and for words to be spoken with pauses between them. However, devices which do not require enrolment or which allow 'connected' speech, are becoming more widely available. Also, commercial devices currently deal with whole words. Researchers have developed algorithms to deal with speech at a lower level, such as phonemes. This may help to improve recognition performance. Therefore, one can expect the capability of speech recognition to improve rapidly in the coming decade.

ii.) Speech output

Speech output from the computer can take one of two forms. Simple speech output can be obtained by using prerecorded human speech. When the output is required, the recording is played. This can give very clear speech output, but is limited by the fact that all messages must first be prerecorded by a human speaker. Such a system will lack the flexibility required for control room operation.

The second form of speech output employs some form of **speech synthesis**. Speech synthesis aims to construct a spoken message given input text and a set of conversion rules. The rules may be derived from human articulation or from signal processing, e.g. analysis of formant frequencies. Current systems use a form of "table look up" to provide information for the pronunciation of spoken output.

Commercial speech synthesisers currently produce robotic sounding speech, but much improvement in speech quality has been achieved in recent years. Speech synthesis could provide a useful means of providing alarm information, or an additional form of information display.

Recognition or Synthesis?

At a very early stage in this project, it was decided that the research should concentrate on ASR rather than speech synthesis or speaker identification. Initially, it was suggested that synthesised speech might be useful in conveying information concerning a particular alarm state. Rather than having a ringing bell, an alarm could 'speak' information concerning the alarm state.

In control rooms, typical alarm situations result in several alarms occuring at once. The initial response of operators is to turn off the alarms and search the visual displays for relevant information. This suggests that if synthesised speech was used for alarms, the following problems could arise:

a.) several synthesised messages could allow confusion of information between messages, or could result in the deterioration of messages into babble;

b.) the operator uses the alarm as a signal to search for information rather than as an information bearer, thus if the alarm was used as an information bearer the operator may have problems in extracting the information from it ;

c.) time is at a premium in alarm situations. If the operator is required to 'decode' the information in the spoken alarm, he might waste valuable seconds. Also, it seems plausible to imagine that until they are confident in using spoken alarms, operators will continue using information from the back panel to check the information in the spoken alarm, thus taking more time.

The problems of speech synthesis, as a means of feedback, are discussed in more detail in chapter eleven.

Speech Technology in a Human Computer System

In order to consider the potential of ASR for human computer interaction, one must place it in a human computer system loop (see figure 1). A human computer interaction loop can be defined by three properties:

i.) the human operator;

ii.) the media of interaction;

iii.) the task.

Figure 1 shows how these properties interact with each other to produce the complex pattern of a working system. Each of the main sections represents a collection of issues which require consideration in this thesis.

The medium of interaction, in this instance, is the voice recognition device. It can be described in terms of a number of factors, according to whether the operator needs to enrol it before use and what type of speech it will accept. These factors will need to be considered in the selection of the ASR device, and the design of the complete system. Of special interest here, is the limitations imposed by the device on the size of the vocabulary, i.e. how many words it can use.

Human operators can be considered from three perspectives:

i.) in terms of their speech characteristics;

ii.) in terms of their information processing capabilities;

iii.) in terms of their workload.

Speech characteristics can be defined by the age and gender of the speaker, and by the way they articulate speech. Most ASR devices will be speaker dependent. This means that, providing the operator speaks in the same manner during both enrolment and use, the device will not be adversely affected by regional accents. However, ASR is unable to deal efficiently with speech impediments, such as stuttering. One way around this problem is to tailor the vocabulary to suit the user's speech. While this is acceptable for applications for disabled people, it does not seem feasible for control room use. The control room system will need to use words and phrases which are already commonly used in the control room, for the identification of objects or for the definition of commands. If operators were allowed to tailor their vocabulary, there is a possibility that they could either forget their 'new' words or substitute them for the accepted control room words. Further, by restricting definition of vocabulary to accepted words, it is easier to provide some system security and reduce operator error.

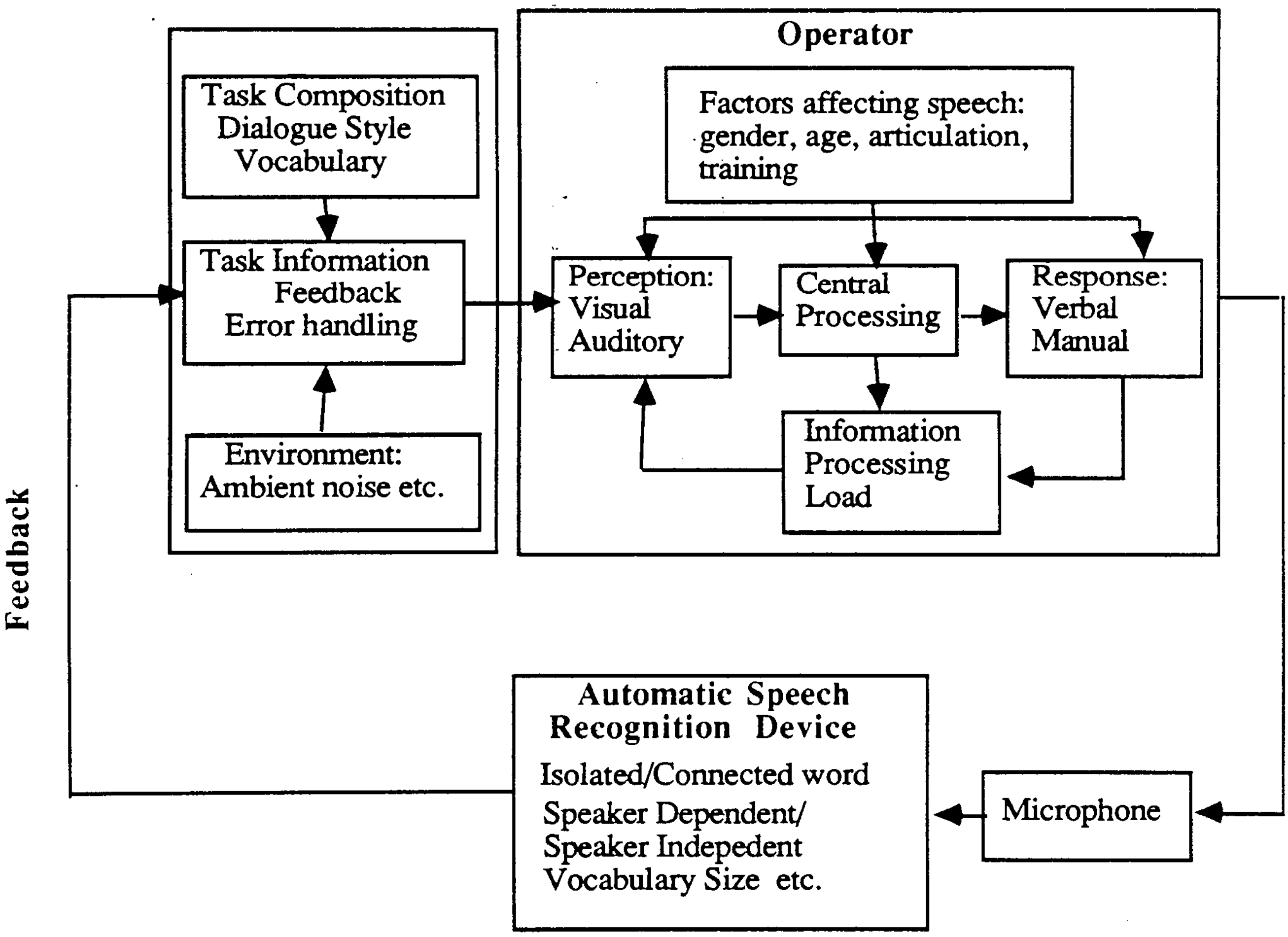


Figure 1: Schematic Diagram of an ASR System
[adapted from Helander et al. 1988]

However, while ASR can deal with consistent speech, it is not easy to define 'consistency' (see chapter thirteen). This means that operators will require some degree of training before they can use the device, and some level of practice before they can reach optimum performance.

Human information processing is the subject of extensive research in cognitive ergonomics. The outline in figure 1 draws on a definition of human information processing which uses three stages: perception; central processing; response. This definition is derived from the work of Christopher Wickens and his colleagues, and is considered in detail in chapter fourteen. Briefly, it proposes that humans process information depending on its style of presentation, e.g. whether it is visual or auditory, whether it is verbal or visual spatial. This processing then leads to the requirement of a specific type of response, e.g. manual or verbal. If the appropriate response is used, then processing compatiblity can be obtained which will enhance operator performance and reduce the likelihood of human error. This factor also contributes to the operators' performance under high workload, which is further investigated in study eleven.

The task can be defined in terms of using ASR and in terms of controlling the process. The issue of task composition relates directly to that of workload, and has already been mentioned. The design of the dialogue between operator and ASR device requires some attention. Obviously, people will not speak to a computerr in the same way that they will speak to other peolpe, but research is required as to exactly how they will conduct dialogues with ASR.

The information operators receive from the system will effect not only their performance, but also how they interprete the ongoing process. This means that adequate feedback is crucial, both for ASR use and process control. Also, ASR will never 100% efficient. This means that the operator will need some form of error detection and correction, in order to use the device to its full potential. ASR use is affected by the environment in which it is used, so careful consideration needs to be given to the problems of ambient noise and other environmental factors.

Both Bussak (1983) and Holmes (1984) suggest that one of the reasons for the small uptake of ASR in industry is that there has not been enough research into the human factors problems associated with its use. This could, of course, be a way of excusing a technology of limited capability, but it does serve to highlight the necessity of fitting ASR to human requirements. This is true of most forms of human factors (Meister, 1989), but is especially true of research in ASR.

System design for ASR should be able to capitalise upon the strengths of ASR use, and not be affected by its weaknesses; it should be able to fit the ASR device to the operators style of working, and could offer ways of improving performance and efficiency; it should consider all aspects of operator performance, and the task domain to produce an efficient system.

General Aims and Outline of Thesis

Although ASR has found numerous applications in industry, it is impossible to predict how well it might perform in the cognitively complex environment of the control room. For this reason, it is necessary to conduct a detailed project into its viablility. Furthermore, there are many questions relating to human factors of ASR use which need to be addressed. From the discussion of "ASR in a human computer system" presented above, five main areas arise:

* Training of both the operator and the device;
* The description of Dialogue for ASR use;
* How to present Feedback to process control operators;
* How to allow operators to correct Errors; and
* How operators will perform with ASR under high Cognitive Workload conditions.

These topics will each receive consideration in this thesis. Before presenting the human factors research, it is important to understand the problems of recognising speech by computers. Further, while different devices use different algorithms, the system designer often needs to be able to compare the performance of these devices for a specific application. Finally, the performance of ASR in comparison with other input media is important, in order to reduce the likelihood of using ASR inappropriately.

Outline of Chapters

Chapter Two: The Human Operator in Process Control

This chapter provides an overview of the role of the human operator in modern process control systems. It considers the nature of operators' tasks and the importance of operators' knowledge in the performance of both routine and nonroutine operations. The discussion of operators' activities in this chapter will inform subsequent decisions concerning the use of ASR in the control room.

Chapter Three: The Recognition of Human Speech by Computer

This chapter discusses the nature of speech, and some of the factors which make ASR difficult. It considers some of the numerous approaches which have been tested in the thirty year history of the subject, and concludes that, although ASR does not perform anywhere near as well as human speech processing, it is a viable means of human computer interaction.

Chapter Four: Technical Aspects of ASR

The aim of this chapter is to provide the reader with a limited, working knowledge of ASR technology. It begins with a brief discussion of human speech production, concentrating on the acoustic aspects of speech. This leads to a consideration of the potential problems of ASR, and some of the more common techniques developed to solve these problems. The reasons for the poor performance of ASR, when compared with human speech use, are explained. It is anticipated that future ASR devices will incorporate more speech knowledge, and some of the earliest knowledge based ASR devices are described.

Chapter Five: Assessing ASR

Before serious consideration can be given to implementing ASR, one must be able to assess how well it will perform. The issues surrounding assessment of ASR are complicated and problematic, as this chapter shows. There are, as yet, no agreed criteria for assessment. However, from a review of previous research and experience in assessing ASR, some proposed guidelines for assessment are provided.

Chapter Six: Current Applications of ASR

This chapter demonstrates that, despite its apparent limitations, ASR has found many applications in areas from office systems to avionics, products for the disabled and telecommunications. The discussion is centred around the use of ASR in industry , as this was initially believed to hold more parallels with control room operation than the other areas. The chapter draws together the considerations from all application areas to provide a set of criteria for application of ASR, primarily in terms of its potential advantages.

Chapter Seven: ASR vs. Manual Computer Control Media

The main medium of computer control in current use is the keyboard. If ASR could be shown to provide better performance than manual control, then it would deserve serious consideration. This chapter investigates the use of ASR and keyboards on 'simple' and 'complex' tasks. Study One compares the use of ASR with function keyboards on a 'complex' task, and shows that task complexity, of itself, is not a sufficient criterion for ASR selection decisions. The main conclusion is that ASR is preferable to manual control in tasks involving 'complex' verbal demands on the operator.

Chapter Eight: ASR in Grid Control Rooms - a Feasibility Study

The recommendations generated in chapters four and five are used to provide selection criteria for ASR in a grid control room. Study Two took the form of a

Hierarchical Task Analysis of operator behaviour. The chapter begins with a short discussion of this methodology and the reasons for its choice in this particular study. Operator activates are briefly described. While accepting that the description may not convey the full complexity of grid control room operations, I would point out that it is being used purely for the selection of possible tasks for ASR application rather than a full description of behaviour. Following this H.T.A., operations are assessed in terms of the possibility of using ASR. The main candidate selected was Telecommand operation. This involved the construction and issuing of commands to sites remote from the control room. Consequently, it can be described as a 'complex' verbal activity.

Chapter Nine: **Design and Evaluation of a Speech based Telecommand Demonstration**

The telecommand task was analysed to provide a rough design for a speech based demonstration. The demonstration simulated the main components of the telecommand activity. In Study Three, the demonstration was assessed with the assistance of grid control room operators. The results were very favourable. ASR offered a viable and effective means of issuing telecommands. Operators felt that it was superior to their existing system.

Chapter Ten: **Speech based Interaction with Computers in the Control Room**

The issue of "dialogue" has been problematic for human computer interaction (H.C.I.) research. The problems are compounded when one is using a 'conversational' medium, such as speech. Before discussing dialogues for ASR, the term 'dialogue' is examined as a metaphor in H.C.I. Its limitations as a metaphor of human computer communication are described, and it is concluded that, rather than basing the design of ASR systems on human 'dialogues', one needs to consider how people will be inclined to talk to a computer.

Study Four shows that people speak very different to other people than to computers, even when performing the same task. Study Five shows that the most prevalent form of speech to computers is short, succinct and task specific. This suggests several points for the design of dialogues. Study Six investigates the role of feedback on user responses. It shows that the more verbose the feedback, the greater the degree of confusion on the part of the user as to what constitutes an appropriate style of speech. This leads to a decrease in user performance.

Chapter Eleven: **Feedback for ASR users**

This chapter investigates the uses of feedback in ASR. A general discussion of the principles surrounding 'feedback' in ergonomics and H.C.I., is followed by a specific discussion of feedback in ASR. Study Seven shows that textual feedback is the most appropriate form of feedback for verbal decisions, such as error detection.
Study Eight shows that if text is provided in a text window, as it will need to be in control room displays, this can lead to an increase in user error. Therefore, a combination of both textual and symbolic feedback is recommended for use of ASR in control room systems.

Chapter Twelve: **Correcting Recognition Errors**

ASR will inevitably make errors. Therefore, it is important to consider how best to deal with such errors. In addition, users will make errors and it is important to discover ways of reducing user error. Drawing on the results of studies presented in earlier chapters, the discussion presents ways in which user error can be reduced. Consideration is given to 'intelligent error' handling for ASR. It is concluded that, although useful, such techniques will still require the user to control error correction. Recommendations are drawn from the studies reported elsewhere in the thesis, together with observations of user performance.

Chapter Thirteen: **User and Device Training**

Despite being claimed as a 'natural' medium for H.C.I., ASR requires some initial training and practice before users can achieve acceptable performance levels. Study Nine shows that a demonstration by an experienced user of ASR can greatly assist learning. Most commercially available ASR devices are speaker dependent.

Study Ten presents brief investigation of the enrolment process, and concludes that users should be allowed to pace enrolment to their own speed, rather than rely on device paced enrolment. Obviously, enrolment is a problem, and ways of reducing the time spent in enrolment are considered. Further, the use of prerecorded templates means that any deviation from 'normal' speech, such as in stressful conditions, will automatically lead to a reduction in ASR performance. Consequently, some means of adapting the recognition process is deemed necessary.

Chapter Fourteen: **The use of ASR in situations of high cognitive workload**

The use of ASR in conditions of high cognitive workload is considered in this chapter. Workload is addressed from the perspective of multiple resource theory of attention. Research supporting this theory is reviewed, after a discussion of attention. Study Eleven reports the use of speech control of a process plant, paired with either a verbal or a spatial secondary task. The verbal secondary task produced the most interference. This suggests that the use of ASR as a medium for command and control in control rooms need not disrupt spatial tasks, such as monitoring.

Chapter Fifteen: **Conclusions**

The studies reported in this thesis are summarised, and their conclusions drawn together in the form of guidelines for the application of ASR in control room systems.

Future research requirements are outlined, together with anticipated benefits of using ASR as a medium for HCI. It is stressed that ASR should be considered not as a novelty or gimmick, but as a possible alternative to conventional input media.

CHAPTER TWO

THE HUMAN OPERATOR IN PROCESS CONTROL

In this chapter, the role of the human operator in the process control room is briefly described. This will provide a framework for consideration of ASR as a means of human computer interaction in the context of control room systems.
The operators' activities are considered in terms of:
tasks, involving data entry and command and control;
knowledge, involving "mental models" of process and plant assumed to be held by operators;
incidents, involving diagnosis and correction of system malfunction and aberation.
The majority of operator activities are concerned with routine operation. For this reason it is proposed that ASR be used for routine operation, especially in the context of command and control.

Intrtoduction

There are many industrial processes which are complex enough to warrant specialised systems of control. The production and generation of electrical power represent examples of such processes. For the human operator, their complexity arises from a number of factors. There are a multitude of inputs to the process under control and the range of variables to be controlled. The process itself is constantly changing, which results in a degree of uncertainty concerning the state specific items. The control of the process carries with it an element of risk, either in terms of plant malfunction or in terms of economic constraints: the operator seeks to keep the process as safe and as economic as possible. As process control strives to be more economic, that is, to yield high production at low cost, operators are forced to make more decisions in the control room during real time operations.

Computers can provide a capacity for centralised information retrieval and display that, in many ways, supersedes traditional knobs and dials control. However, there are very few situations in which operators will rely solely on the computer for data. They will also use the mimic displays around the control room walls, and information recorded on paper or passed over the telephone.

The introduction of computers produced a change in the role, and work, of the operators. Prior to computerisation, operators were mainly responsible for recording and analysing data. This required large amounts of paperwork to be completed, and imposed heavy demands on operators' time and attention. While computerisation has reduced this 'clerical' aspect of operators' work, it has led to an increase in other aspects. In the modern control room, activity is of a more analytical nature. For instance, in the grid control centre, studied in chapter eight, a computer scans the power stations for performance measures. Operators make judgments and decisions based on this data and their knowledge of the ongoing process.

Many aspects of industrial process control have been automated; it could be argued that, providing no adverse effects impinge on the process, the entire plant could be computerised. However, a wholly computerised plant would not be realistically feasible. It would be impossible to design a plant that would be able to cope with all the potential faults that can occur in process control.

Bainbridge (1987) argues that some plant designers will automate the 'easy' parts of the process first, and then any other parts that can be automated. This leaves the operator with a set of 'left over' tasks, which have very little consistency or coherence. Rather than exploiting the operators' skills, such systems tend to relegate the operator to an adjunct to the computer.

Knowledge based systems are being investigated for process control (see Hollnagel et al, 1988), but it is difficult to emulate the human operators' creativity in problem solving or to cpature the whole of their knowledge. This means that the operator needs to be retained as an essential part of the control loop.

Computer based control and display technologies can significantly enhance operator performance of routine tasks, but they can also impose heavy information processing demands upon the operators. In order to assess the magnitude of such problems, it is necessary to have at least an idea of what constitutes process control operators work. The following section presents a review of process control tasks. This is followed by a consideration of the role and use of operator knowledge in the design control room systems.

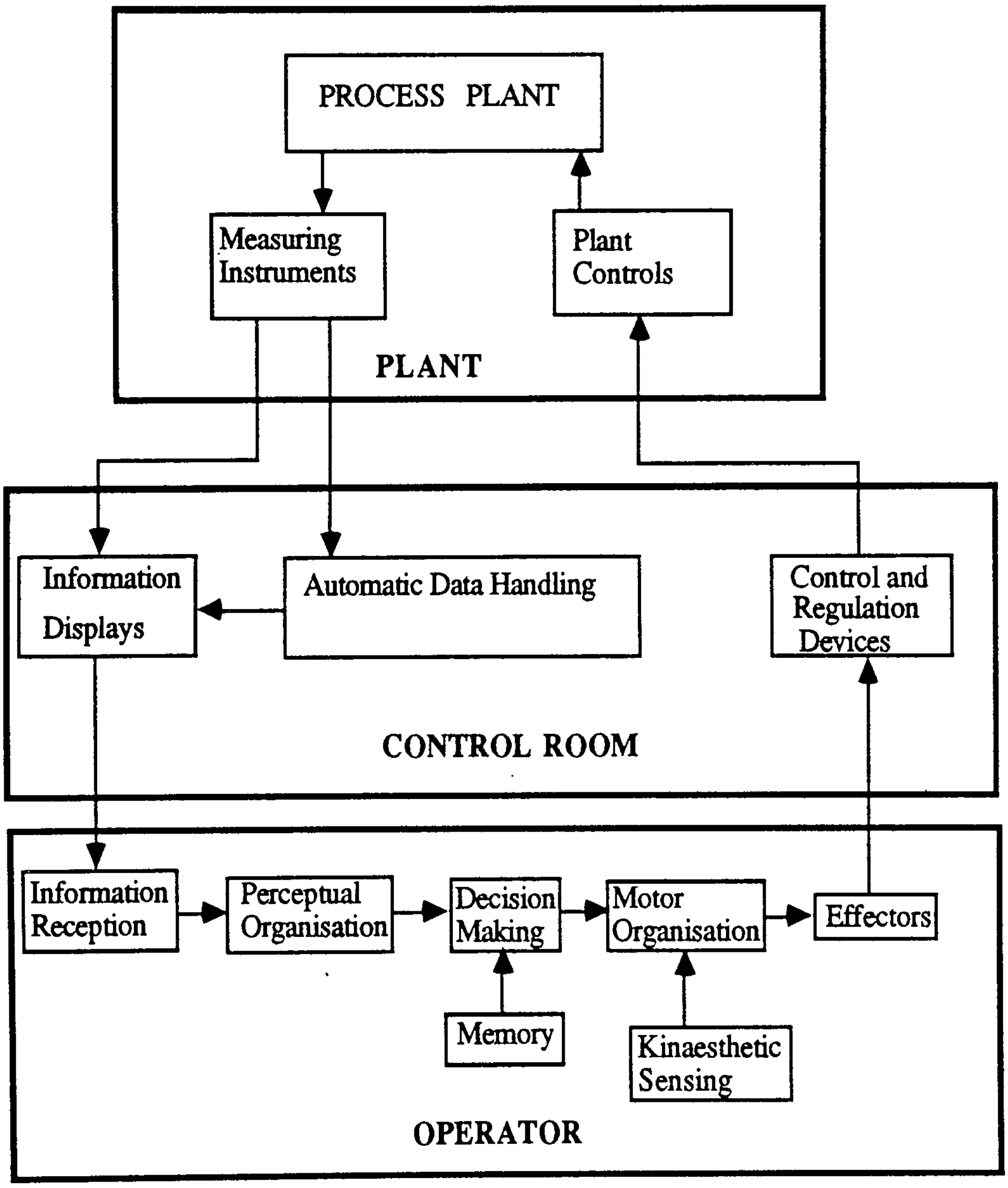


Figure 2.1: Overview of Process Control Systems

Figure 2.1 shows the relationship of the human operator to the process under control. Although this is only meant to be a simple schematic diagram, it shows one important aspect. The control room acts as a transducer between operator and plant. The operator does not directly control any aspect of plant activity, but enters data or commands into computers in the control room, which are then transformed into system operations. This is important to note because the function of the operator is not one of direct controller of the plant. That is, when the operator alters the position of a particular knob on a control panel it will change a particular system state. However, the control is not direct, but an analogue of direct control. This point is clarified when one considers that many of the operations in a modrn control room are carried out by typing instructions to a computer, rather in the form of altering physical controls.

Tasks of the Control Room Operator

There has been growing interest in the human factors study of control room operations in the last thirty years, although the U.S. National Research Council (1983) recognises that there is still much needed research to be carried out in this field. Despite marked shifts in technological capabilities in the control room, operators' tasks are based on a relatively constant set of requirements and goals. The changes in work introduced by new technology tend to alter the ways in which these requirements and goals are met.

At the most basic level of description, the human operator could be defined as an intermittent correction servomechanism (Craik, 1947). That is, the operator forms part of a control loop and seeks to optimise the process. In a simple control loop, this optimisation could be defined as the maintenance of process output within a specified tolerance around a specified target level.

While this is sufficient to describe control of a single variable, the multitude of variables in a particular process mean that the definition is far too limited (Bainbridge, 1981). Further, while the control of a single process can generally be performed directly by the operator, process control operation is carried out with the operator at least one remove from the process. Operation is performed in a central control room, with data concerning all plant states being received and displayed. The operator then acts in terms of these data.

The performance of control actions comprises a very small part of operators' activites. Umbers (1979) estimated that in one control room, control actions occured, on average, 0.7 times per hour. A cursory look around any control room might convince the observer that control room operators spend most of their time drinking tea and doing very little! Basing a description upon physical behaviour hides the vast proportion of control room tasks. In addition to the low number of physical actions performed, there is a limit to the time scale over which actions can be performed. Many of the process variables have extremely long time constants (Wickens, 1984). This means that, although a cntrol action takes the form of a discrete change on the control panel, the process will respond in an analogue fashion. Thus, the operator needs to keep a check on the results of a particular control action for some time after it has been performed.

The first point to note, in defining control room activity, is that can be divided into two basic types: routine and incident. Routine operations are concerned with the well planned, stable running of plant. Incidents represent situations in which performance falls below the optimum level and can result in some cost, usually defined economically but cost can also be considered in terms of safety. The majority incidents result in the incurrance of some economic cost. From interviews carried out as part of Study Two, operators can be assumed to spend around 90% of their time carrying out routine operations, and 10% dealing with incidents. If the process under control could be perfectly predicted and modelled, then operators would not need to deal with incidents. However, such incidents occur from a number of unforeseeable causes. The following discussion concentrates on routine activities.

We have already noted that operators perform control actions. and that they need to check the results of these actions. The control actions, in turn, are based on the operators' assessment of the current state of the process.

Thus, the process control operator can be seen to perform three basic tasks, taken from Bauerschmidt and Le Porte (1976):

i.) extract data from the environment (or process);

ii.) manipulate these data to provide specifications for plant control actions;

iii.) communicate these control actions to the process via the operators controls.

(This tripartite definition is also supported by Rasmussen (1974), who offers the terms: measuring function; data processing; control action).

The extraction of data from the process can be performed by direct measurement, either by the operator or by staff at the plant level, or, more commonly, by monitoring plant displays. To date, most process control rooms are constructed around the principle of "one measurement - one indicator" (Goodstein, 1981). This results in physically large control rooms, with panels containing hard wired meters and other forms of display. Operators are faced with the daunting task of processing information from the most basic level of analysis, before they can begin to make decisions concerning plant and process states (Rasmussen, 1983). This inevitably leads to an increase on the possibliltiy of human error, as the following quotation from a special edition of Human Factors, illustrates,

> "Large volumes of data are often presented at rapid rates and in different formats, forcing the operators to select, integrate, and interpret information from various sources. The demands of such multitask environments may often exceed human information processing capabilities, thereby increasing the probability of error and performance breakdowns."

[Preface to Human Factors. 30(5). 1989]

This problem has traditionally been addressed with the use of numerous alarms for various process behaviours. Such an approach is not without its own problems, placing an extra task on the operator in the interpretation of alarms. The lack of any adequate definition of what constitutes an alarm, only serves to confuse issues further (Stanton and Booth, 1990). Current research is aimed at producing an intelligent means of presenting plant and process performance data to the operator via computers (Hollnagel et al, 1988).

The manipulation of these data, requires the analysis of the information in the displays to derive not only current process state, but also to specify appropriate control actions. In many control rooms, such actions are, in turn, defined in written procedures. However, in order to perform the actions efficiently, the operators need to posses far greater knowledge than the procedures tend to assume. This knowledge will be discussed in the section below.

Manipulation of data is also necessary for the prediction of future states. This latter is important to monitor the growth and development of process trends, which provide vital clues concerning the possiblility of process malfunction.

From the observations of grid control operators, reported in chapter eight, data processing took several forms. Operators would input data directly into the computer, either to complete performance tables or to give commands. Operators exchanged information verbally, either over the telephone (to national control, to power stations, or to other grid control rooms), or to colleagues. Operators logged data and operations by hand.

Control actions are communicated to the plant via the input and control devices in the control room. These controls were originally knobs, switches and buttons hard wired onto control panels. However, computerisation has led to these being replaced. This has resulted in the centralisation of control operation. Rather than having to walk around the panel, altering some controls and leaving others, the operator is able to enter control actions from a workstation. In general, the workstation is equipped with a keyboard for command and data entry. However, research by Carey (1985) has suggested that the keyboard may not be the most efficient type of input medium for all control plant operations (this point is discussed in more detail in chapter seven).

Control actions can either take the form of continuous tracking control or sequential controlling (Edwards, 1976). That is, the operator can either alter a changing variable, for example, using a control wheel or a joystick, or can control a changing sequence. The latter will normally involve the setting of targets for the process to reach over specified times.

Kragt and Landeweerd (1974) interviewed process control operators to ascertain their opinions of the most important aspects of their tasks. Operators tended to emphasis four points: monitoring the process, effecting changes in the process in response to disturbances; minimising the effects of process or plant breakdown; handing over control of the process. Of these, we have already dealt with the first two, and the third will be discussed below.

Changing over control of the process is very important in most control operations, which run on shift cycles. One could define change over in very broad terms to be assuming control of the process. This would then include the important tasks relating to changes in the operational state of the plant, such as plant start up and shut down. Start up and shut down are largely the function of automatic controls, with the operator monitoring their activities and only intervening to begin the start up or in an incident.

When a shift change occurs, and a new operator takes over the process, it is common for the oncoming operator to 'sit in' and watch the outgoing operator. This allows the operators to exchange information concerning any problems in operation on the previous shift, any processes still in progrees and any anticipated changes or problems. It also allows the incoming operator to develop a 'feel' for the current state of the plant and process.

Operator Knowledge in the Control Room

Bainbridge (1981) notes that while the performance of physical tasks in the control room is minimal, the operator is required to perform several cognitive tasks. A major activity is the setting of goals or targets for operator, process, or plant performance. These goals can take four forms:

i.) Developing goals into operational form

Often the goals received from central control or the management are not fully specified. This is generally because the goals are derived from projected performance figures, which cannot take into account all the variables which will affect the process. The operator has to define the goal in terms of this range of variables, as they are effecting the process at the time of operation.

ii.) Predicting targets from varying product demand or process behaviour

For instance, electricity cannot be stored in large quantities. This means that if it is not used when it is produced, it will be lost. In order to maintain an efficient and cost effective electricity generation and distribution system, it is important schedule the production to coincide with demand. This scheduling needs to be as accurate as possible, and so is revised on a regular basis.

iii.) Plannning sequences of future behaviour

This is partly a result of the activity of scheduling, but also involves coordinating the various processes under control, each with different time constraints. It also covers the aspect of process prediction, mentioned above.

iv.) Generating new plans for novel situations.

The operator needs to deal efficiently with incidents, by formulating appropriate plans of action (see below).

It has been shown that operators are able to control of process with only a very limited amount of knowledge (Paternotte, 1978). In these circumstances, control is exerted by small, increments in the intended direction, and only requires knowledge of the direction and a very approximate idea of system dynamics, i.e. gain of control movement. Such control actions are typical of situations in which operators are dealing with processes with a high degree of uncertainty. This would suggest that some limitation in the overall design makes it difficult for the operator to develop sufficient knowledge to formulate a more efficient strategy (Bainbridge, 1981).

One of the most common activites, which operators draw on knowledge of the process to perform, is fault diagnosis. In a study of fault finding in a simulated process plant, Marshall et al, (1977) found that operators tend to use two distinct approaches.

The first approach was, contrary to their hypothesis, the use of alarm data to diagnose faults rather than the scanning of analogue values according to diagnostic rules. It seemed that the operators mistrusted process indications, and prefered to wait for confirmatory evidence in the form of an alarm before searching for faults. This phenomenon, of amassing confirmatory evidence before taking action, is well supported in the literature on behaviour in hazardous situations (see Hale and Glendon, 1987).

This approach led to problems when the fault gave rise to alarms in different areas of the plant. Furthermore, research by Woods (1984, 1988) shows that while operators are efficient at spotting faults in the early stages of a process, they tend to make mistakes as the process continues. He assumes that this results from a decoupling of the operators perception of plant state from the actual state. This leads operators to take actions corresponding to their idea of what is happening, rather than coupling control action to process response.

The second approach found by Marshall et al, (1977) occurs when the general area of the fault was identified. On isolating the area of the fault, the operators began to use systematic procedures and heuristics in diagnosis.

These procedures, however, tend to be general rather than specific to each type of fault. Rasmussen and Jensen (1974) show that electrical maintenance engineers tend to rely on procedures appropriate for many instruments rather than investigating each instrument in terms of its function and working.

Bainbridge (1984) argues that fault finding relies on both knowledge of the variables in the process and the performance of individual components. From this, she concludes that diagnosis takes two forms, "aggregation", in which individual process variables and global functional knowledge is used (see Marshall et al, 1977; Rasmussen and Jensen, 1974), and "abstraction", in which knowledge concerning the physical components and abstract entites is employed.

This suggest that the operators knowledge of the plant and process is built from several different sources. These are combined into quite simple procedures for general working, often in terms of cause and effect propositions about plant behaviour (Cooke, 1965). Thus, initially knowledge of the process takes the form of propositions concerning a correlation between states and actions. This type of simple knowledge is sufficient for routine operation. In such situations, the operator may even be happy to hand over control to a layman as he knows that only limited knowledge is necessary (Kragt and Landeweerd, 1974). If additional information is given to the operator in these situations, there is a tendency for performance to be inhibited; too much information leads to a misinterpretation of the process state (Kragt and Landeweerd, 1974).

There has been much debate concerning how these knowledge types are represented internally in the operator. The most common definition relies on the term 'mental model'. Mental models can be defined as internal, mental representations of a process or system held by the operator (Crossman, 1956; Bainbridge, 1974). The literature on internal representation of information proposes two information codes: verbal or visual (Norman and Rumelhart, 1975; Wickens, 1984; Paivio, 1986), although this is not without its critics (see Pylyshyn, 1986). This distinction, between verbal and visual information codes, is developed by Landeweerd (1979) to cover process control models. The verbal, internal representation refers to the functioning of the process, in terms of "what leads to what" production rules. The visual, internal representation refers to the structure of the process components, for example, in the form of a plant diagram.

This basic distinction can be further supported from the work of de Kleer and Brown (1983) defining several types of mental model. Of interest to this discussion, are the notions of "device" and "attribute" typologies. The "device typology" is a structured schematic of system components, and their interconnections; roughly speaking it is a type of 'mental map' (see Pick and Acredolo, 1983, for a discussion of the topic of 'mental maps'). Naturally, one would expect such information to be either stored or at least represented in visual terms. Landeweerd (1979) found that operators were able to draw a diagram of the system they had been controlling, with reasonable accuracy.

The "attribute typology", on the other hand, contains process values for the interconnections between components, in terms of values for general concepts such as pressure, temperature, flow. Landeweerd (1979) found that operators were able to accurately answer questions, such as "if x is open, what happens to Y". Thus, on tasks requiring different aspects of the process to be considered, operators rely on different forms of internal representation of the information.

From interviews and task analysis, three major types of knowledge concerning process control have been defined (Allengry, 1987; Hoc, 1987). These are:

i.) Knowledge concerning the current state of the plant and process (usually in terms of measured values);

ii.) Knowledge of the structure and interrelationship of plant components;

iii.) Knowledge of the developing process (in terms of observed and predicted performance trends).

While the internal representations described above clearly contain information concerning plant structure and performance, they do not describe the developing process (Coury and Pietras, 1989). This means that planning and predicting future operation or assessing the risk of potential faults cannot be performed using these representations alone. Rather, an additional type is needed, such as the underlying rules according to which the process develops (Moray and Rotenberg, 1986).

Crossman (1956) proposed that operators use their knowledge in three distinct ways. Routine operation can be dealt with by 'rules of thumb', derived from experience or procedures. Or it can be dealt with using a 'mental model', which represents an accumulation of information into an intuitive appreciation of the workings of the process, that is, a 'feel' for the process. In general, experienced operators tend to use the latter type of knowledge, while less experienced operators use the 'rules of thumb' which can be readily acquired.

In incident situations, the operators need to engage in more complex logical reasoning. This distinction between three levels of operator knowledge bears a strong similarity to that proposed by Rasmussen (1983) and termed, skill based, rule based and knowledge based.

It is clear that the 'rules of thumb' Crossman (1956) speaks of relate to Rasmussen's (1983) concept of 'rule based' knoweldge, and that complexlogical reasoning relates to 'knowledge based' knowledge. The relationship between 'skill based' knoweldge and 'process feel' is ,at first glance, less clear. However, it should be appreciated that 'process feel' often operates at a level below consciousness; it constitutes an intuitive reaction to process operation, rather than being overtly conceptual. This is very similar to many skilled behaviours. It is highly efficient, automatic (in the sense that it requires minimal overt cognitive processing), and not fully available to conscious introspection (this latter point is ilustrated in research using verbal protocols, see Bainbridge, 1974 and Ainsworth and Whitfield, 1983).

This brief discussion illustrates the potential complexity of analysing the knowledge possessed by control room operators, and shows the wealth of experience they need to develop to perform efficiently. To summarise the types of knowledge defined, I will borrow four terms from Rasmussen (1979). This allows the muddle of terms and definitions used in the preceding discussion to be condensed into an unambiguous set.

i.) Physical Models: such as mimic diagrams or plant displays;

ii.)Functional Models: comprising plant descriptions in terms of similar objects, e.g. boilers, pumps, or in terms of functionsal relations, e.g. feedback loops, together with rules specifying relationships.

iii.) State Models: (not discussed above) comprising 'snapshots' of the plant in different states, e.g. start up, power failure. These can be drawn upon for fault diagnosis and problem solving.

iv.) Behavioural Models: representing the dynamic process in operation, and can be used to predict trends and future events.

This discussion is intended to illustrate the point that operators do not have a single 'mental model', rather they have a complex of representations and knowledge concerning process and plant performance. Ideally, a system designer should be able to provide the right type of information, in the right type of format to operators for different types of decision and information processing task. This point is developed in the studies reported in chapters nine and twelve.

DEALING WITH INCIDENTS

We have briefly investigated a few of the concepts associated with routine operation in the process control room. While this can explain around 90% of operations, it cannot explain the remaining 10%. These operations are concerned with fault handling and incident recovery. While it is conceivable that routine work could be carried out by expert systems

(Hollnagel et al, 1988), it is not possible to replace the experience, expertise and creativity embodied in control room operators by a computer.

Incidents, by their very nature are nonroutine and, often, unpredictable. This means that operators are unable to rely on routine skills, but must develop new ones. It is possible that the incidnet may have been encountered before by the operator or by a colleague. In this instance, the procedure which worked before is employed. Occassionally, operators will telephone staff at other sites to draw on their expertise and knowldge, especially if it is known that the other sites has experienced similar problems.

Incidents will either be dealt with by the individual operator, or by a group of operators and supervisors. Generally, the operator needs to draw on his knowledge of the current state of the process and the preceding states in order to evaluate the incident. From here, he can begin a problem solving procedure to isolate the fault further, until it can be discovered. Then the appropriate action can be taken. There has been a wealth of work researchnig into the problem of coping with incidents in process control, (e.g. Edwards and Lees, 1974; Rasmussen and Rouse, 1979; Woods et al, 1987; Sheridan, 1988; or Moray and Rotenberg, 1989).

CONCLUSIONS

From this discussion of the human operator in process control, it is possible to note that different types of operation will require different types of information. For changes in operational state, such as start up and shut down, the operator needs to be kept informed of changes in the process and plant states. For routine operation, the operator will rely on primary information concerning the major aspects of the process. For nonroutine incidents, the operator will need a clear presentation of current measured values and process behaviour. Additionally, it has been suggested recently that he will also benefit from expert advice in the form of knoweldge based systems to assist his problem solving (Holnagel et al, 1988). It is possible to present information to the operator via synthesised speech. This could reduce some of the overloading in terms of visual processing. Obviously not all types of infomation can be presented using synthesised speech. nor can it be used in all situations. Clearly this is a topic which needs research.

In addition to different forms of information presentation, there is a need to research different types of control device (Carey, 1985). New technologies are being developed which might allow easier interaction with computer systems. As interaction becomes easier, the task of using a computer will become less of a burden and allow the operator to concentrate on primary tasks. Using human speech to control the computer and as a means of data entry could offer benefits to the operator, particularly when he is performing two tasks simultaneously, such as reading data from a display and entering it into a computer. This is topic will be investigated in this thesis.

CHAPTER THREE

THE RECOGNITION OF HUMAN SPEECH BY COMPUTERS

This chapter presents an overview of the nature of human speech, and some of the problems associated with designing computer systems which can deal with speech. A review of the various approaches which have been tried in the thirty year history of the subject is presented.
It is suggested that, although very limited when compared with human speech processing capabilities, ASR is able to perform well enough to be considered for certain applications. Some of the more recent developments are discussed, and it is proposed that the technology of speech recognition is developing rapidly.

The Nature of Speech

In constructing and speaking a series of words, a speaker must make a number of decisions. Initially, the information to be communicated must be planned and selected in terms of the appropriate linguistic and social contexts of the conversation. Obviously words cannot be spoken in a random order, but must adhere to a recognisable order, defined by the syntax of the language.

Two influential theories concerning how humans produce speech have recently been proposed. Garrett (1976; 1984) begins with an analysis of the type of error people make in producing speech. From this analysis, he proposes a four stage model of speech production. Dell (1986) approaches the problem of speech production from the viewpoint of 'spreading activation' theory. In this theory, activity in the brain is believed to spread to other areas, producing corresponding activation, until speech is produced. Dell (1986) also proposes a four stage model.

The models of both Garrett (1976; 1984) and Dell (1986) are sufficiently similar for them to be conflated into a single, four stage model. At the highest level, the meaning of the message to be communicated is generated. This semantic or message level, uses the speakers knowledge of language and the context of the conversation to construct a basic intention to speak and the content of the speech. This obviously has to be translated into speech.

At the next level, the basic outline of the message is generated. This syntactic, or functional level, corresponds to the setting up of a row of 'slots' which need to be filled by certain types of words (although highly simplistic ,this example illustrates the activity at this level).

Following the syntactic level, comes the morphological or positional level. At this level, the words selected to fill the 'slots' are constructed from the appropriate morphemes (basic units of meaning) of the language.

Finally, an articulatory program needs to be developed and run to produce the speech. At each level, a set of rules need to be invoked in order to produce coherent speech. This would suggest that, as neither Garrett's (1976; 1984) nor Dell's (1986) model have a 'control' level, more research is needed on the basic phenomena of human speech production.

These models illustrate that a number of constraints operate on the speaker to produce coherent speech, as figure 3.1 shows. Without delving into the various linguistic arguments concerning speech production and generation, it is possible to note from this diagram that speech is made up of several units. These units are combined according to certain rules to form larger, meaningful units (Nolan, 1986).

Once the message has been received by a listener, the process described in figure 3.1 needs to be carried out in reverse. There are, of course, differences between producing and understanding speech, but it is conceivable that these activities share similar types of processing (Clark and Clark, 1977). At the highest level of processing, the listener will need to use knowledge shared with the user, concerning the world they live in and the language they use. Without such knowledge it would be very difficult to make any decisions about the meaning of the speech.

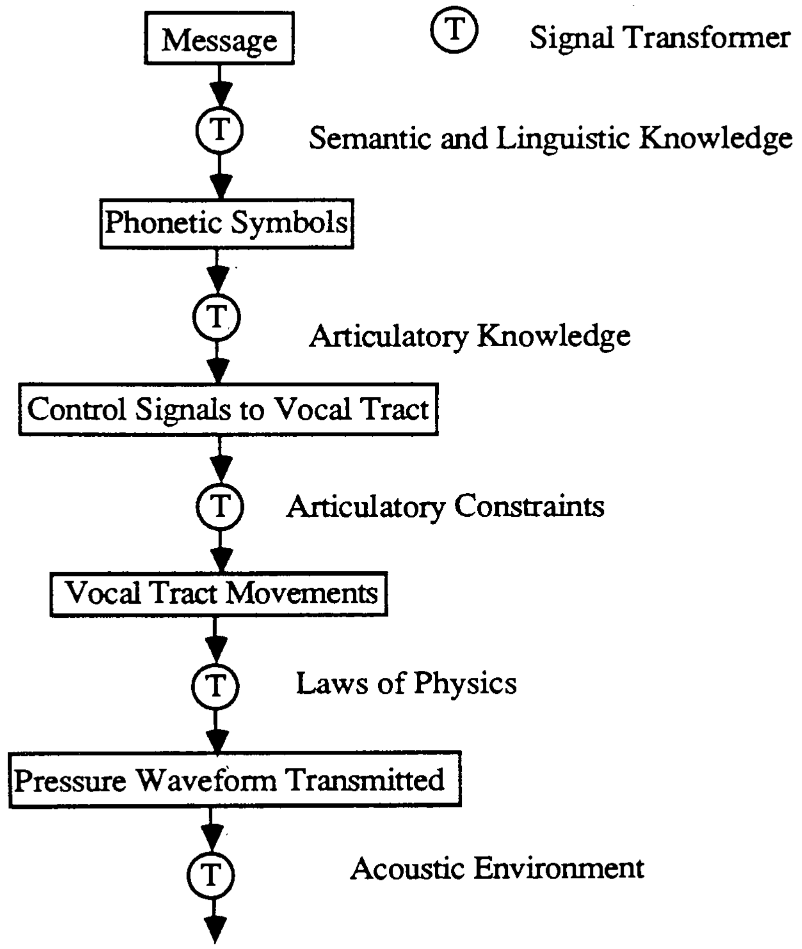


Figure 3.1: Stages of Processing for a Spoken Message
[from Underwood, 1977]

However, it is possible to decide whether the speech is in one's own language or not, and whether the speaker is male or female. Such decisions would only require information from the acoustic waveform. Chomsky and Miller (1963) propose that certain acoustic changes, such as those related to emotion or fatigue, are redundant to the process of recognising words.

This does not mean that such changes are not useful paralinguistic cues; they can signal different meanings of the same word spoken in different contexts (Halliday, 1973). It does, however, suggest that one can extract information from the acoustic signal, and that one can limit the amount of information required to recognise speech at different levels of interpretation. If decisions based solely on the acoustic waveform could be used for the 'lexical' interpretation of speech (deciding what word has been spoken), then conventional signal processing techniques could be applied to speech recognition devices.

Unfortunately, speech is far more complicated than the types of waveform generally dealt with in signal processing, but it is possible to achieve some degree of success using such limited techniques. Lea (1986) suggests that ASR research can been viewed from four distinct perspectives:

i.) Speech Production - using theories developed from human speech production;

ii.) Acoustic - using signal processing techniques;

iii.) Sensory Reception - using models of the human ear;

iv.) Perceptual - using linguistic information to inform the recognition process.

Perspectives (i), (iii), and (iv) can be defined as "knowledge... driven representations" (Newell, 1975). This means that they require information beyond that contained in the speech signal in order to carry out speech recognition. As the following discussion will show, by far the most common technique in use is perspective (ii). Although the acoustic techniques were devised in electrical engineering, they bear some resemblance to the 'source filter' theory of speech production, familiar to linguists, and the discussion of ASR will begin with a consideration of human speech production. This will be followed by a review of some of the techniques used in ASR, and a discussion of knowledge based ASR. Perspective (iii) receives no consideration, as it is a relatively untested approach (although see Chistovich, 1979; Fujisaki et al. 1986; Pols and Plomp, 1986).

Human Speech Production

Human speech may be defined, in acoustic terms, as a time varying sound wave originating in a vocal mechanism which is in constant movement during phonation. We learn to speak during our early years of life. Such learning suggests that speech is a skill, which requires coordinated movements to be performed on the air in the respiratory and vocal tract. These movements must contain some degree of consistency for recognisable speech sounds to be produced (Pickett, 1980).

The production of speech sounds can be explained using the source-filter theory (Müller, 1848; Fant, 1960; Liberman and Blumstein, 1988). In this theory, the larynx acts as the sound source for speech. The waveform generated in the larynx is passed through the supralaryngeal vocal tract (pharynx, mouth, teeth, tongue, lips etc.) which acts as an acoustic filter.

The airflow from the lungs is modulated by the vocal cords in the larynx. This acts as the sound source in speech production. The vocal cords open and close very rapidly during the phonation of voiced speech, causing a quasi periodic waveform to be generated. The rate at which the vocal cords open and close determines the fundamental frequency of the voice which results in variations in the pitch of the voice.

Despite this constant change, the components of the tone produced in the larynx are always harmonics of the fundamental frequency. When the vocal tract resonates, a peak is produced in the sound spectrum closest to these harmonics, which ensures that the resulting spectrum always has the same envelope, even though the fundamental frequency is continually changing. This is very important because it allows a "sameness of quality" in the range of sounds around the fundamental frequency (Fry, 1979).

During speech production the dimensions of the filter vary because the vocal tract is in constant movement. Different shapes of the vocal tract can be considered as filters with different transfer characteristics: allowing maximum energy through at some frequencies while suppressing energy at others. The frequencies at which maximum energy passes are the formants. In general, although higher formants do exist, people tend to use only the first three formants in the spectrum.

In everyday terminology, we tend to divide speech sounds into two classes: vowels and consonants. Linguists generally describe vowels as vocoids, which are the simplest type of speech sound to sound describe in acoustic terms. Vocoids are those sounds produced when there is no obstruction in the vocal tract to impede the airflow. Rather, the airflow is modulated in the larynx and by the position of the tongue.

Matthei and Roeper (1983) define vowels as steady state concentrations of energy at certain points in the speech spectrum. They can be distinguished by their different first and second formant frequencies, with relative ease. These formant frequencies relate to differences in tongue position associated with their articulation, as shown in the 'vowel quadrilateral' in figure 3.2.

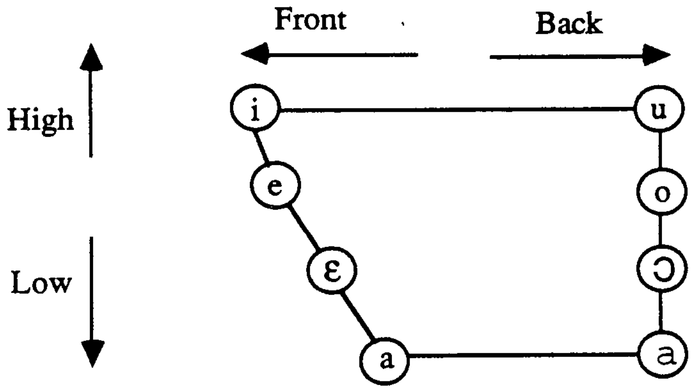


Figure 3.2: The IPA Vowel "Quadrilateral"
[From Liberman and Blumstein, 1988]

When there is an obstruction in the vocal tract, a contoid is produced. Contoids correspond to what are usually called consonants. Figure 3.3 shows the relationship between the position of the obstruction and the consonant produced. The distinction between vocoids and contoids is informative because it relates the production of different types of speech sound to the appropriate areas of the speech production apparatus.

TYPE OF CONSONANT		PLACES OF ARTICULATION									
		bilabial	labio-dental	dental	palatal	alveolar	postalveolar	velar	uvular	pharyngeal	glottal
STOPS	voiced	p				t		k			
	voiceless	b				d		g			
AFFRICATES	voiced						tʃ				
	voiceless						dʒ				
FRICATIVES	voiced		f	θ	s	ʃ					h
	voiceless		v	ð	z	ʒ					
APPROXIMATES	central	w					r	j	(w)		
	lateral				l						
NASALS		m			n				ŋ		

Figure 3.3: Table of Types of Consonant and their Places of Articulation

[From Nolan, 1986]

There are two main differences between vowels and consonants. The first is in type of sound source. Vowels always produce periodic waveforms, whereas consonants could result in periodic, aperiodic, or a combination of both types of waveform. The second lies in the fact that aperiodic waveforms result from constrictions in the vocal tract. Thus, consonants require shaping of the vocal tract to a greater extent than vowels. Stop consonants, / b,p,t,d,k,g / close the vocal tract, obstructing the breath stream, completely before releasing it in a short burst. Fricatives, such as / f,v,s,z/, are formed with narrow constrictions. Nasals are formed with a constriction of the passage between mouth and nose. Finally, the 'approximates' (see figure 3.3) are midway between consonants and vowels, and are often called semivowels.

The different points of constriction, and type of waveform, result in different formant frequencies. In theory, therefore, every speech sound can be associated with a characteristic set of frequencies. If this were the case then ASR would simply be a matter of pairing the incoming frequencies against a set of reference sounds, one for each type of speech sound.

Unfortunately, ASR is not nearly so simple. The speech signal is highly dynamic and effects such as coarticulation (the running together of speech sounds in normal speech) together with the variations in frequencies for speech sounds across speakers and even for the same speaker in different situations, means that speech waveform is very complex to process.

The Problems of ASR

Human speech processing is a highly complicated affair. As Fry (1979) points out, the acoustic information only provides the "rough scaffolding" for the reconstruction of the speaker's message which comprehension entails. This process of reconstruction is carried out using our knowledge of the language and processes of inference generation and testing. However, this knowledge based processing relies strongly on an ability to correctly recognise speech sounds and assign them to their correct phonemic categories.

ASR is basically a process of information transformation. Information presented in one domain of representation (human speech) must be transformed into another (a numerical representation of the speech waveform). The speech signal needs to be captured in workable samples, which need to be analysed in order to create a stored representation. This stored representation can then be used as the basis for the recognition of incoming speech. The incoming speech can be matched against the stored representation. Therefore, on a very simple level, the stored representation could be said to contain the 'knowledge' necessary for the recognition of speech. A device which could carry out this 'matching' is shown in figure 3.4.

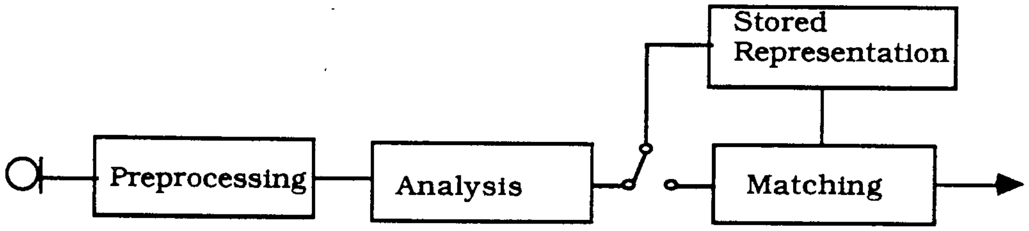


Figure 3.4.: Schematic Diagram of a Simple ASR Device

This would suggest that speech recognition could successfully be performed by implementing algorithms which transform the acoustic signal received by a microphone into "useful parameters" (Fallside, 1985). The initial acoustic signal would be defined as a time varying waveform, rather than as speech. Consequently, it would not be necessary for the algorithm to have any speech knowledge in order to process the signal. As we shall see in later sections, this approach has met with some success in commercial ASR systems.

In order to reduce the complexity of the recognition task, researchers initially adopt certain simplifying assumptions. In any stretch of speech it is assumed that each articulation of a specific word will take a constant time, i.e. words are assumed to always be spoken at the same speed, and that words are spoken in isolation, i.e. the beginnings and endings of words will be easy to detect and will not overlap with preceding/ following words. Such assumptions are not justifiable in terms of normal human speech, but are of much value technically.

Commercially available ASR devices are a long way from even coming close the matching human speech recognition ability. Future research into speech understanding devices might help some of the problems.

A Brief History of ASR

The idea of controlling machines by using one's voice has appealed to people for centuries, but it is only since the 1950's that such an idea has stepped out of the realms of science fiction. The past 40 years have seen some significant developments in the technology of Automatic Speech Recognition (ASR), which allows users to control a wide range of systems in simple tasks. Therefore, it is now possible for us to emulate the ability of Dr. Dolittle and talk, not to animals but, to computers.

i.) Early Attempts

The idea of being able to command a machine by using one's voice has intrigued scientists for centuries. The engineers of the Seventeenth and Eighteenth Century Automata sought in vain to incorporate voice commands into their devices. Reddy (1980), points out that "speech recognition is an interdisciplinary problem that has it's roots in centuries of historical studies of language, sound, physiology and automata". It was not until the Twentieth Century that a combination of knowledge in all these areas was sufficiently developed to allow a serious study of ASR to begin.

It was becoming increasingly obvious, to scientists at the turn of the Century, that the speech signal was a sound wave whose amplitude varied as a function of time, and which could be decomposed into simple sinusoidal waveforms, using signal processing techniques such as Fourier analysis. If speech could be broken down into simple components, then, it was suggested, a machine could be built which used these simple components to 'recognise' the speech signal.

The first step in this process would be to convert the speech signal into a form which the machine could deal with. It is possible to convert speech into an electrical signal using a microphone or a telephone, but in order for a machine to begin recognition of speech it will be necessary to convert the signal to a machine readable form.

Flowers (1916) proposed using speech as an input to telegraph transmission. At that time, the received telegraph signal was converted to a paper tape output which a trained operator could read. Flowers suggested that rather than transmitting the message as an 'alphabetic code', it would be possible for a user to speak the message into a transmitter. The signal would then be passed to the receiver, where it would be broken down into twelve frequency bands.

Energy in each frequency band drove an electromagnet, which in turn moved a tiny mirror. The mirror reflected light onto a photoelectric cell, the current of which drove a recording pen. The paper output was in a 'phonographic code' which an operator translated. Flower's (1916) paper does not describe any performance figures, but this is probably the first reported attempt to mechanically convert speech from signal to paper.

In the 1930's and 40's, the increase in radio technology led to the development of the 'sound spectrograph' by researchers at Bell Laboratories (see Koenig et al. 1946). This has proved to be the most successful attempt at capturing speech as a graphic representation. The principles behind speech spectrographs are discussed in chapter four.

One could argue that the task of a speech recognition device would be to read the 'graphic representation' and convert it to specifiable features in the speech signal. These features could then be used to 'reconstruct' the spoken input. Exactly how the speech signal is captured, how the 'features' are defined, and how the process of matching features with speech signal is performed remain the central topics of ASR research.

Reddy (1980), bestows the honour of being the "earliest known speech recogniser" upon a toy dog called Radio Rex (date unknown). Strictly speaking, Radio Rex was a voice activated device rather than a speech recogniser, i.e. it did not require specific words to be recognised so much as certain sounds. When one spoke his name, Rex would jump out of his kennel. However, Rex's performance was affected by two factors which are still of major concern in the field of ASR today.

The first being that he could be activated by a large number of noises which only slightly resembled the word "Rex". The problem is to restrict the definition of the function word to a single word, but to allow several speakers to use the device. Each speaker may pronounce the word slightly differently so the definition needs to be able to cope with these differences without accepting erroneous words.

The second problem related to the quality of the user's voice. Any major changes would greatly affect performance. The problem is to provide some compensation for variation in the voice due to fatigue, illness etc. without drastically altering the templates or overloading the memory. Incidently, it is interesting to note that this very early application of ASR was in a toy, considering the huge expenditure of toy manufacturers into ASR today (see Viglione, 1986).

One potential application that has been suggested for ASR since its conception, is that of the "phonetic typewriter". This is a machine that can automatically transcribe a spoken utterance as a piece of typewritten text. Poulton (1983) reports an incident concerning a German scientist named Nemes. In 1930 Nemes applied for a patent for a 'phonetic typewriter'. But the German patent authorities rejected the idea as being "impossible in principle". There have been several speech typewriters reported, but they tend to work on very small vocabularies and with limited success (see Meisel, 1986)

Another early attempt at speech recognition was a device invented by Dreyfus Graf (1950). In his 'stenosonograph', a speech signal was analysed by a band of six filters. Each filter represented an essential resonance of the French language. The filtered signals were then passed to deflection coils arranged in a circle, to produce a deflected electron beam on a cathode ray screen. The beams position depended on the relative energies of the six filters. But a human reader was still needed to decipher the display.

ii.) 1950-1960 : Early Successes

It was clear from the use of speech spectrographs, that different words produce different patterns. This observation led to the development of one of the first ASR systems, which was described in a paper by Davis, Biddulph and Balashek (1952). (It should be noted that several other institutions were working on similar projects at around the same time). This paper described a system which could automatically recognise and distinguish the digits from zero to nine. The system used a principle which is still common today.

The input signal was recognised by comparing it against a set of stored 'templates' to select the best fit. The 'templates' were constructed from a simple analysis of the acoustic waveform for particualr words. The user spoke each word several times, and an averaged 'template' was produced. This process of repeating each of the words in the vocabulary is known as enrolment, and is still in common use in speaker dependent ASR devices.

Figure 3.5 shows a schematic diagram of this automatic digit recogniser. The incoming signal is divided into two frequency bands (<900Hz. and >900Hz.). The number of times the signal in each band passed through zero volts was counted to provide a value for 'axis-crossings'. The signals were then passed to the axes of a display to give what is now termed a 'formant1- formant 2' plot. This pattern could then be cross correlated with patterns already stored to provide a best fit recognition. The hardware was kept simple by limiting the vocabulary to ten words (zero (oh) to nine), and by not providing a facility for rejecting illegal words. After extensive enrolment by a single speaker, the system could recognise digits with an accuracy of between 97 and 99%.

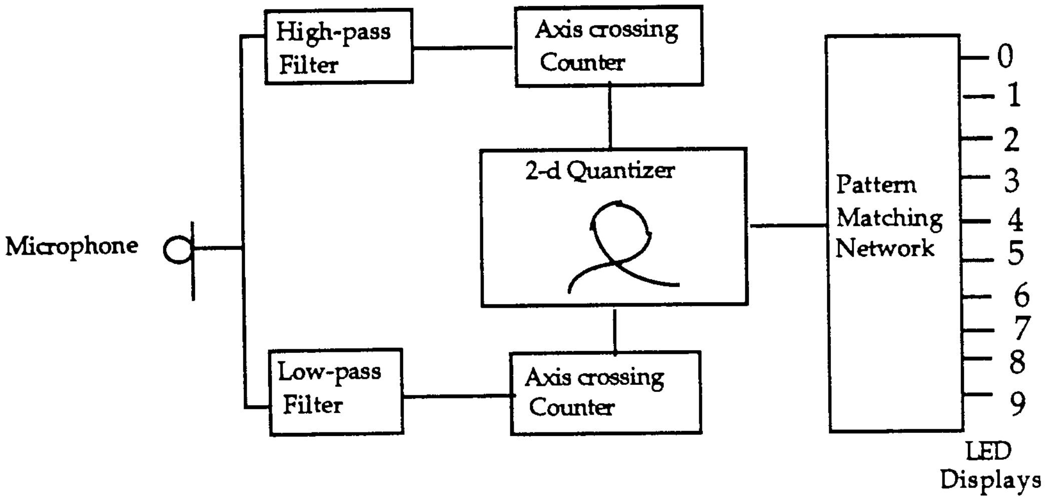


Figure 3.5: Schematic Diagram of Davis et al. (1952) Digit Recogniser

An improvement on the system of Davis et al (1952) was developed by Dudley and Balashek (1958). Their system, "Audrey", could recognise a limited number of stored word patterns spoken by a single speaker. The speech signal was divided into ten frequency bands, and "Audrey" derived certain spectral features from these signals to compare with stored feature patterns. Words were segmented into phonetic units which could be identified from their spectral patterns. Recognition accuracy of 99%+ was reported for a single speaker, once the system had been trained. If a new speaker tried to use previous templates, accuracy dropped significantly. Problems with considering Audrey as a commercial system were its immense size and slow computing time.

As Ainsworth (1988) points out, the early ASR devices rely on the acoustic speech signal; it was felt that all the information necessary for recognition was contained here. The close of the 1950's saw a dawning realisation that the simple matchnig of acoustic patterns would only work on very limited speech recognition problems.

iii.) 1960-1970: Further Developments

Developments which took place during the 1960's, included attempts to produce small marketable products, e.g. Dertsch (1961) "suitcase recogniser"; and work in developing recognisers to work in languages other than English. Researchers were also trying to expand the size of the vocabularies that systems could efficiently use. Gold (1966) managed to obtain a recognition accuracy of 86% with a 54 word vocabulary. Bobrow and Klatt (1968) managed 97% with a vocabulary also of 54 words. However, the late 1950s and early 1960s represent a quiet period of ASR research. Although the systems developed so far were speaker- dependent, there was still a problem with variation of input words. One solution to this problem came with the one of the earliest applications of digital computers to ASR. Digital computers also allowed research into ASR to proceed apace.

Denes and Matthews (1960) developed a spoken digit recogniser on an IBM 704. Utterances were converted to time/frequency patterns, which were correlated to test patterns for each digit. Utterances were reduced to standard durations. This technique of time normalisation is an important contribution to ASR. By bringing all utterances to a standard duration, i.e. slowing down fast speech and drawing out slow speech, variation in speech rate could be compensated for before recognition occurred. Using a group of five male speakers, Denes and Matthews (1960) obtained scores of 6% error with time-normalisation, 12% without.

Time normalisation is now frequently done by a process known as time warping (Itakura, 1978), which requires the application of dynamic programming techniques which were introduced in the early 1970s.

iv.) 1970-1980: Microchip technology and new approaches

The 1970s saw a growing realisation that ASR was more than simply a matter of sophisticated pattern recognition. Research was conducted into providing ASR devices with linguistic and 'world' knowledge (Vicens, 1969; the ARPA SUR project of 1971-1975). Although research showed that knowledge based ASR could work, systems are still in the early stages of development and will not be commercially available for some time.

Techniques for processing speech signals as dynamic waveforms were developed and improved during the 1970s. Linear Predictive Coding (Itakura, 1975) has been employed with some effect, and led the way to developments in dynamic programming techniques. Further, the speech signal was assumed to vary in line with statistically predictable probablities. Markov modelling was used to determine these probablities and generate templates and matching algorithms to exploit them. Markov modelling is the currently the most popular technique in research system.

In commercial systems, advances have been made in silicon ASR devices (see Quarnby, 1986), allowing recognition power to be concentrated in ever smaller products. However, in general, the devices do not utilise novel techniques but rely on performing old techniques in smaller spaces. Finally, there has been an increasing interest in the possibility of phoneme based ASR devices (see the OSPREY project), and it is expected that ASR research will follow these lines for some time. Also. the possible use of parallel distributed processing equipment for ASR is receiving attention (Vaughan et al, 1987; Tattersall et al, 1988).

In conclusion, there is much research being performed into the development of better ASR devices. However, commercial products lag some way behind research products. This makes human factors research problematic.

One can either use existing technology, and run the risk of any findings being quickly outdated, or one can use a simulation of a perfect ASR device, in the form of a human listener, and run the risk of being criticised for being too optimistic concerning the power of ASR.

I have decided to carry out studies using both approaches, but to concentrate on the latter. This will allow the device dependent problems to be ruled out of experiments. and reduce contamination of results, and will allow basic principles of human behaviour with ASR devices to be uncovered and studied. Ultimately, this will lead to the development of system design guidelines for optimum ASR system performance.

CHAPTER FOUR

TECHNICAL ASPECTS OF AUTOMATIC SPEECH RECOGNITION

ASR is very primitive when compared to human speech perception. This chapter presents an outline of speech production, and the techniques used by ASR to capture and recognise speech. The discussion is based on the source filter theory of human speech production and concentrates mainly on the acoustic aspects of speech production. It is the acoustic signal, from the microphone, which provides the primary information for ASR. It is argued that the acoustic signal is far more complicated that ASR designers had first realised, and that in order to design efficient devices, it is necessary to incorporate a great deal more linguistic knowledge into the recognition process. Despite its obvious limitations, however, current ASR capability does offer some potential for human computer interface design.

Commercially Available ASR Devices

Speech Technology (April/ May, 1989) provided a buyers' guide to currently available speech technology devices. Thirty eight ASR devices were listed. They were classified as speaker dependent or independent, isolated or connected or continuous word devices, and according to vocabulary size. Figure 4.1 shows a classification scheme of ASR devices.

IWR refers to Isolated Word Recognition. This describes the simplest type of ASR device, which can only recognise a single word or short phrase at a time. The user is required to pause between each item of vocabulary, until the device has recognised what has been spoken.

CWR refers to Connected Word Recognition. Such a device is capable of analysing a string of words spoken together, but not normal speech rates.

CSR refers to Connected Speech Recognition; a system which can handle normal, conversational speech.

Talker ch.c refers to the type of system which uses the characteristics of a person's speech, building samples of such characteristics, to allow recognition of items which have not been enrolled.

Talker adaptive refers to the ability of the ASR device to adapt its recognition process to variations in the users' speech.

Figure 4.1: A Family Tree of ASR Devices

[From Sedgwick, 1987]

ASR devices can also be characterised according to whether they need to be trained to recognise an individual's speech. The collection of speech samples is known as enrolment. Devices which require enrolment are known as speaker dependent, in that they depend on the speaker providing adequate speech samples before they can be used. Devices which do not require user enrolment are known as speaker independent.

In the Speech Technology review, twenty two speaker dependent devices were reported, compared with sixteen speaker independent devices. This shows the wider availability of speaker dependent devices in the current market. As speaker dependent devices require enrolment, which can be time consuming, one might be tempted to conclude that a speaker independent device would be preferable.

However, the speaker independent devices listed in the guide show an average vocabulary size of fourteen to twenty words. These words are pretrained and supposed to be common to almost all applications. In control rooms, a degree of adaptability is essential and commands must be specified for specific plant operations. Therefore, one would need to use a speaker dependent device.

In a review of other ASR guides, shown in figure 4.2, the distribution of device types noted in Speech Technology is found elsewhere, with isolated word recognisers being most common. However, there is a noticeable increase in the availability of connected word recognition devices, and of devices with larger vocabularies.

			SPEAKER DEPENDENT	SPEAKER INDEPENDENT
SOURCE	NO. X IWR	NO. X CWR	NO. X VOCAB.	NO. X VOCAB.
Peacock and Graf (1990) [12 devices]	5	7	8 (64- 20k)	4 (80- 30k)
Speech Technology (1989) [38 devices]	22	13	22 (64-1k)	16 (14-40)
Wallich (1987) [17 devices]	12	5	15 (64-1k)	2 (13-40)
Lea (1983) [13 devices]	13	0	10 (40-500)	3 (10-40)

Figure 4.2: Commercially Available ASR devices

This review gives an idea of the availability of ASR technology, and illustrates some of its limitations. Devices are capable of dealing with a small to medium size vocabulary, which generally needs to be enrolled before use. There are reports of systems employing vocabularies of several thousands of words, but there is little evidence that such devices are capable of fully exploiting their vocabularies. As we shall see in chapter five, the accuracy of an ASR device tends to deteriorate in proportion to the number of words in the vocabulary. Therefore, it is reasonable to assume a working vocabulary of between 64 and 1 000 words. Considering the number of words an average speaker of the English language knows, (circa 20 000), this might seem too small to be of any use. The question of vocabulary size is addressed in chapters eight. The next section will discuss some of the underlying technical theory of ASR, in an effort to explore the complexity of the speech recognition problem.

Approaches to ASR

The problems of ASR can be attacked using one of three separate approaches. Basic signal processing techniques can be employed to give efficient recognition performance on very small vocabularies of easily distinguishable items. Alternatively, pattern matching techniques which were introduced into speech recognition research as early as 1952, have shown to provide highly robust performance, superior to signal processing alone.

During the development of these techniques several important ideas have been tested, and template based pattern matching is held to work as well as it ever will. It provides a computationally cheap and efficient method of analyzing short chunks of speech within vocabularies of up to several hundred words. An improvement on matching templates relies on stored, statistical models of speech.

The third approach adopts ideas and concepts from the field of artificial intelligence, and can be regarded as a knowledge based approach. It is, as yet, untried in commercial applications and exists only in research models. However, the central premise behind this approach is that human speech recognition relies on several sources of information for its efficiency. In order to mimic human speech recognition ability, it is necessary to incorporate as many sources of information as possible.

It is possible to draw parallels between these approaches and theories proposed to explain how people recognise visual patterns. The earliest theories of pattern recognition in perception, proposed that people matched objects in the world with stored representations in their brain. These either took the form of exact replicas of objects, or prototypes (typical examples of a particular category of concept). Recognition was thought to involve matching object with stored image. Problems occur, with this approach, when trying to explain how, when typical objects are slightly altered, e.g. viewed from an unusual angle, they can still be recognised. This can be dealt with by assuming a process of normalisation, in which stored representations are altered to fit the perceived object. This is possible, providing one is not confronted with novel objects for which there are no stored representation.

A second approach suggests that stored representations are constructed from a number of features (Selfridge, 1959; Morton, 1969). Recognition then involves activating a number of features. When sufficient features have been activated, the object is recognised. There is a growing body of evidence from neurophysiology to support this proposal (Lettvin et al., 1959; Hubel and Weisel, 1962, 1979; Grossberg, 1988).

The main difference between the two approaches is one of grain of analysis. Where the former uses a coarse representation of the objects concerned, the latter analyses at a deeper level in terms of features. However, it could be argued that these approaches are still limited in that recognition, to some extent, involves knowing what has been recognised. Therefore, as well as representing objects, some form of decision process is also involved in perception. People construct their perceptions from often fragmentary evidence, by calling upon assumptions drawn from previous experience and knowledge of the world.

This extremely brief discussion outlines a number of points which are developed in the ensuing discussion. It is worth noting that, in comparison to research on perception, ASR is largely still at the contentious 'stored representation' stage rather than considering the finer grain of feature level analysis.

The stored representation can take the form of a template, constructed from several enrolment utterances by a single user, or a statistical model of salient features of the words, derived from very many utterances from many speakers. Whichever approach is used, it is necessary to normalise the incoming speech with the stored representation. This can be carried out be using some form of time alignment, and will enable the computation of measure of similarity between spoken word and stored representation (Mc.Innes and Jack, 1988).

Pattern matching techniques

Before creating reference patterns (templates), some sort of reduction in the overall data in the speech signal needs to be carried out. This is necessary in order to provide the ASR device with enough information to carry out the analysis, but not too much information as to make the task impossible. Early devices used such characteristics of the signal as zero crossing rate (Davis et al. 1952). This captured enough information for simple discrimination of very small vocabularies, but was not suitable for larger vocabularies.

The simplest method of template matching is to represent words individually, in a crude fashion, as a set of features derived from a filter bank analysis of the speech signal. The 'distance' (or difference in energy level at different frequencies) at successive points in each set, for corresponding parts of the word, can be compared. The sum of the 'distances' can then be calculated for the whole word.

Speech is sampled at a preset rate and analysed using a bank of filters. The energy levels at each frequency are then recorded at regular intervals. This information is stored in the form of a numerical representation, or template. Matching is carried out by comparing the representation of the incoming pattern with the stored templates until a best fit is found. In order for the templates to be meaningful, isolated word recognition is used (isolated word in this context also encompasses short digit strings and short phrases). In order to calculate how best to space the filters over a suitable frequency range one needs to define an optimum bandwidth over which the device can operate.

The typical fundamental frequency of adult male speech is around 200~300Hz. Voiced speech reaches an amplitude peak at around 800Hz, and unvoiced speech at 4-5kHz. Flanagan (1972) points out that most of the meaningful components of speech can be held to lie below 7kHz, which is why telephone systems can be band limited to around 5kHz without losing too much speech information. Generally, speech recognizers operate in the region of 300Hz- 5kHz. Filters are spaced over this frequency range in a fashion analogous to the signal processing characteristics of the human ear. As frequency increases, the resolving power of the ear decreases. This allows filters to be placed linearly at low frequencies, and logarithmically at higher frequencies. The first operations of the ASR device is thus to remove high frequencies, and to measure the energy of the signal in each of the frequency bands the filters cover.

As well as varying between samples, the speech signal varies within samples. The rate of variation can be assumed to be proportional to the speed at which the articulators (i.e. components of the vocal tract) change from articulating one phoneme to the next. If the signal is sampled at 10ms. intervals, one can assume that the speech signal will be static over these short time intervals. Using 'windows' of 10ms. allows for the tracking of variations in resonances of the speech signal, which in turn can provide information concerning the movements of articulators necessary to determine phonetic information. At the end of each 10ms period, the energy in each frequency is measured and recorded.

This method assumes that phonations of the same word are of exactly the same length, and that variations in energy are of equal importance and vary constantly with each articulation of the word. Unfortunately, such assumptions cannot be justified. Current template matching techniques are completely deterministic. Although they work reasonably well for small vocabularies, they cannot cope with vocabularies of several thousand words. Klatt (1980) states that the calculated scores between phonemes, derived from such techniques, do not compare very well with perceptual scores of human listeners. This suggests that different sources of information are being used by humans to those of the ASR device.

Endpoint Detection

It is important to determine where a word begins and ends in order to analyse it. In optimal conditions (high signal to noise ratios), the start of an utterance can be characterised as a high level of energy, and the end of the word can be marked by a fall in energy level below a certain value. However, speech does not always allow for optimum conditions.

Words beginning with a weak, unvoiced phoneme, /f/, could have this initial phoneme confused for noise. Words beginning with a strong vowel are occasionally produced with slight lip opening noises. At the end of words, unvoiced stop consonants can be reduced to 'flaps', or omitted completely. Vowels can vary in length between different articulations, and between different consonants. It is impossible to determine whether low level sounds are part of a word or noise, without determining the identity of the word.

Stop consonants are preceded by periods of silence, which can vary in length between different articulations of the same word. Generally, ASR devices require periods of around 100-200 ms. to elapse before they accept that a word has terminated. This is easy for isolated word recognition, where users are required to pause between each word they speak.

It has been assumed that words will be of a constant length, and that frames in the speech signal samples will be static to allow frame by frame analysis. However, as has already been pointed out, words vary in length each time they are articulated. This means that some form of compensation is required.

The simplest form of compensation requires time alignment between incoming pattern and stored template. Vowels generally exhibit greater variation in length than consonants, and it would appear that the most useful information in the speech signal can be portrayed by the transitions between phonemes.

Spectral time derivative methods of time alignment, analyse incoming patterns and templates in terms of their rate of change relative to a preset value. In this manner, portions of rapid change can be highlighted. Such portions often correspond to consonants.

Dynamic time warping techniques were introduced to the field of ASR in the 1960s, and are rapidly becoming very popular. The time axis of the incoming signal is nonuniformly distorted or warped, to bring its features into line with the template pattern, in a matrix. A predefined distance metric is used to calculate the distance, d(i,j), between frame i of the incoming pattern and frame j of the template pattern, as figure 4.4 illustrates.

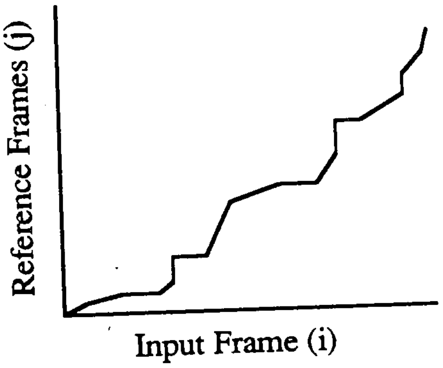


Figure 4.4: Diagram to Illustrate Dynamic Time Warping.

The horizontal steps represent one reference template matching two input frames; and vertical steps represent two reference templates matching one input frame. The overall distance score is calculated by totalling the individual distances between template and input frames, with some weighting per frame to compensate for variations in slope of step. Assuming that different representations of, say "a", will give lower distances than, say between "a" and "p", one can initiate a search through the matrix. The shortest route from bottom left to top right corners can be found using dynamic programming techniques. A best match will result when the shortest route is found.

Dynamic time warping reduces the problems of variations in speaking rate, but introduces problems by removing the temporal relations between successive frames. This information might not be relevant for consonant detection, but words which vary primarily on vowel length, e.g. "bid" and "bit", will be confused as information concerning the duration of the vowels will be lost.

Linear Predictive Coding

Filter bank analysis makes no assumption concerning how the speech signal was produced, nor does it seriously take into account any aspects of speech as a dynamic sound wave. "Time series analysis" techniques have been developed and used in many forms of social science and engineering. Basically, they permit a mathematical analysis of the behaviours of dynamic systems (Makhoul, 1975). Linear predictive coding (LPC) is a version of such analysis.

LPC takes the basic concept of the source filter theory of speech production, and adds the assumption that the vocal tract can be modelled as an all-pole, time varying, linear filter (Mc.Candless, 1974). This produces an idealised model of the vocal tract, and allows the pitch, formant frequencies of speech, and the vocal tract area parameters to be modelled (Smith and Sambur, 1980). The filter is excited by either a periodic sound source (in the case of voiced sounds, such as vowels and nasals), or by noise (in the case of unvoiced sounds, such as fricatives).

LPC assumes that variations over time in the vocal tract can be approximated by a succession of static shapes. Each continuous time signal is sampled a number of times to give a discrete time signal. It is possible to determine the transfer function of this all-pole model, based on the number of poles measured and the input excitation (pitch period). The resulting static shapes can be characterised as a series of different filter effects (Atal, 1985). Thus, each speech sample is represented as a linear combination of its past values and the current value of the input (Makhoul,1975). LPC is a simple and efficient method of representing the short time spectrum of the speech signal.

If the model can be seen to have passed through certain previous states, given the input excitation, then it be assumed to be passing towards a state x which thus can be predicted. This allows the speech signal to be tracked as a dynamic soundwave of varying frequencies, rather than captured as a static representation as in filter bank analysis. So words which are defined purely by vowel length, can be captured, as the duration is shown.

The spectral peaks provide information which is useful in the estimation of periodic sounds, e.g. vowels. But LPC tends to over estimate the bandwidth of sounds which cannot be easily modelled by an all-pole filter, e.g. nasals, which are produced by an irregular filter model of the vocal tract, and fricatives, which are produced by white noise (Ainsworth, 1988). Work by Markel and Gray (1976) produced algorithms to reduce the error signal (the difference between the predicted and the actual signal). Combining filters with LPC, can give information relating to both vowels and consonants to define words.

Statistical pattern matching

The techniques described above make the assumptions that either the speech signal can be captured as a single, static template (filter bank analysis) or that, at some level, it can be characterised as a set of momentarily static positions of the articulators, linked by transitions from one position to the next (LPC). As has been suggested, this does not give a true representation. Speech is essentially a highly dynamic process which involves many coordinated articulatory processes. This means that we need to examine more of the time varying parameters of the speech signal in order to provide a fuller acoustic description. One way of doing this is to model the vocal tract as a continually moving sound source (Jelinek, 1976).

The statistical probability of the vocal tract being in position x to produce sound y can be computed using the stochastic methods of hidden markov modelling (HMM). We assume that the way speech varies is highly structured, and that a direct relationship exists between the speech pattern and the speech production mechanism. Given the speech pattern, we can calculate the position the speech production mechanism was most likely to have been in to produce the pattern (Cox, 1988).

In order to do this, we assume that the speech production mechanism is capable of being in a finite number of states, and that each of these states is capable of producing a finite number of outputs. The transitions between states describe the evolution of the speech pattern over time, thus providing information that we lose in dynamic time warping.

Speech patterns are analysed as a sequence of short time frames, and recognition is carried out by matching incoming patterns with stored representations, as in the other pattern matching techniques described above. However, in HMM the stored patterns are made up of stochastic finite state models. Each time the model for a particular word is activated, it produces a set of speech parameter vectors which represent an example of that word. The best match is the model deemed most likely to produce the incoming speech pattern.

Phoneme Based ASR

An alternative to the holistic pattern matching techniques described above, and one which is receiving much attention in the 1980s, is to use explicit phonetic information, derived from the knowledge of expert phoneticians. Some pattern matching devices are capable of operating a simple form of syntax which helps to constrain the search space, but the recognition process still relies on the matching of numerical representations of the acoustic signals for whole words. We have seen that the use of whole words is useful in that it allows the capture of data for complete patterns for storage, providing adequate space is left between words in use.

Whole word pattern matching techniques are disadvantageous in that they cannot cope with local variation in the structure of the speech signal. Researchers are currently looking at ASR devices which use smaller units of recognition than whole words, e.g. phonemes. It is assumed that smaller units will be less prone to variation in their production. This could be very useful approach to ASR, but the pattern matching techniques would need to take account of additional information. Phonemes are not the simple acoustic sound patterns that engineers would like them to be, as Pickett (1980) points out,

> "It is important to understand that phonemes are determined by their function in differentiating the words of a language. Thus a phoneme is defined within a language system not acoustically."

If we examine different variants of the same phoneme (allophones), we can see that they are acoustically distinct, e.g. the /s/ in "sill", "still", and "spill". In all /s/ has over 100 allophones. Phoneticians define the basic speech sounds, phones, as groups of common, persistent features in allophones of the same phoneme. This would suggest that phoneme based ASR would be prone to similar problems of feature detection as word based. Further, the problem of meaningfully segmenting phonemes to form words would be very difficult without some higher level, linguistic knowledge to guide the process. Finally, these techniques have yet to be implemented in commercial systems.

It is surprising to read that researchers into ASR have failed to explicitly consider the performance of the expert speech recogniser, the human speech user. The section below introduces some of the points which could usefully be incorporated in future ASR research.

Human speech perception

The techniques described above are representative of what Lea (1986) calls 'ignorance' models of ASR. They do not use linguistic information to assist in the recognition process, even though some of the parameters they use can be justified in terms of speech acoustics. One of the solutions to these problems is to mimic human speech perception, to some extent. Although the study of human speech perception is still rapidly expanding, there are several points which are of interest. Humans tend to use as much information as possible in order to decode the speech signal.

Marslen-Wilson (1987) has shown that human speech perception combines bottom up (phoneme tracking) with top down (hypothesis testing) approaches. ASR loses some of the information in the speech signal by losing prosodic cues, and environmental cues. But if we examine the performance of people working with a very limited representation of speech, we can discover what rules and techniques might be useful in ASR. This is the approach favoured by researchers examining spectrograms.

Decoding spectrographs

Several research projects in ASR were triggered by the study of human decoding of spectrograms. In these studies, experts decoded representations of the speech signal, i.e. all the information they were assumed to have was contained in the spectrogram. This meant that such information as is provided by the environment and context in which the speech was produced was lost.

There have been several studies in which an 'expert' is asked to read a spectrogram, and his performance is compared with that of a panel of subjects who listen to the speech signal under test, e.g Klatt and Stevens (1973), Cole et al. (1980). In one study, the 'expert' could identify fluent speech from a spectrogram with an accuracy of around 85%, after several thousand hours of practice (Cole et al.,1980). More recently there have been studies into the development of an expert system to read spectrograms, e.g. Zue and Lamel (1986), Stem, Eskenzi, and Memmi (1986).

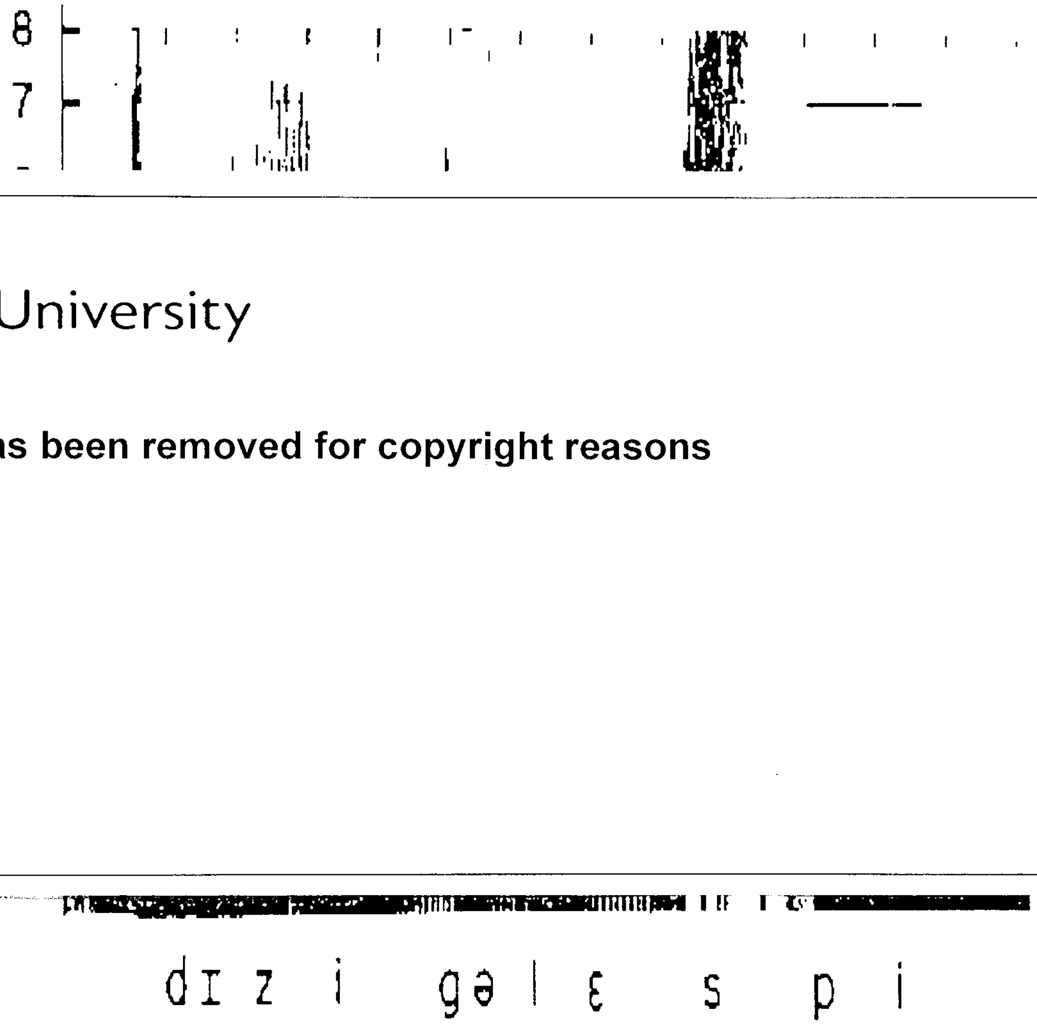


Figure 4.5: A Spectrograph of the phrase "Dizzy Gillespie"
[From Matthei and Roeper, 1983]

It is possible to discern some of the phonemes spoken in the phrase in figure 4.5. The vertical axis represents increasing soundwave frequencies, i.e. the higher the pitch, the greater the the value of y; the horizontal axis represents the passage of time, and is read from left to right. The intensity of the sound (loudness) is shown by the darkness of the mark: the darker the mark, the louder the sound.

From the performance of 'expert' spectrograph readers, it is suggested that the acoustic signal may be regarded as the primary information bearer for speech. If phonetic segments are to be perceived directly from the information on the spectrograph, then these phonetic segments must be accompanied by specifiable acoustic features.

The first task in speech recognition would then seem to be the detection of discrete acoustic phonetic events in the speech signal, and to use such events to 'label' the segments. Experts pick out obvious events, such as high frequency fricatives, and break the spectrogram into recognisable sections of events. If one looks at the spectrogram included here, even a nonexpert can recognised distinct patterns. Next experts apply combination rules, based on their knowledge of the language to combine key points into a coherent message. Paradoxically, this approach stresses the importance of seeing relations between the acoustic components of speech, emphasising the speech signal as an integrated set of acoustic patterns in time, whereas the spectrogram displays segregated frequency patterns

The task of a speech recognition device would then be to discover the specifiable features in the speech signal and use them to 'reconstruct' the spoken input. The study of the methods used by 'expert' spectrogram readers suggests that the recognition task is improved by using linguistic constraints, such as syntax and semantics.

Also, it appears that 'experts' use an identifiable control strategy in determining words from the complexity of spectrogram patterns. This strategy could be implemented into a recognition system to improve accuracy. For instance, it is clear that humans often use information from the context in which speech occurs to make sense of it, e.g. by using the information which has already been presented to interpret forthcoming information and attempt to reduce confusion due to ambiguity etc. However, what this simple theory does not explain is how *linguistic* information is represented in the acoustic signal.

Knowledge Based ASR

It might strike readers as strange that pattern matching techniques of ASR do not use linguistic information in the process of recognising speech. We have examined various techniques, and seen that the major problems of ASR concern the inherent variability of human speech and the difficulty in accurately detecting the boundaries of words.

Moore (1984) has pointed out how early recognizers adopted traditional pattern matching techniques, with the assumption that speech was a highly redundant signal. This would mean that a great deal of information in the speech signal could be ignored. It was hypothesised that speech a sequence of invariant information bearing elements called 'phonemes' (analogous to letters in the written language). As we have seen this view of phonemes is largely discredited by linguists, but it led to the assumption in ASR research that one could discover which words had been spoken simply by looking up the sequences of recognized phonemes in some form of "pronouncing dictionary".

A series of articles by Lindgren (1965) characterised the feeling, during the early to late '60s, that research was being limited by these simplistic assumptions. Researchers were realising that higher level linguistic information was necessary to enable the development of adequate ASR systems. In other words, speech could only be recognized if it was, in some sense, understood. This would require the ASR device to contain some knowledge about the language it was using and the task domain in which it was being used.

Fry and Denes (1953) had incorporated a way of assessing the probability of one word following another in a certain limited vocabulary command language; Wiren and Stubbs (1956) used the theory of 'distinctive features' (Jakobson et al., 1952) to create a binary tree recognition process; Bobrow and Klatt (1968) used simple 'prosodic features' whose 'on- off ' transitions were used to select which words from the lexicon to compare with the input word.

During the late 1960's researchers were trying to develop systems which could cope with "normal" speech, i.e. continuous speech using normal pronunciation and vocabulary. It seems that an unspoken assumption at the time was that continuous speech could be recognised by applying the techniques of individual word recognition in real time. It was soon found that one of the major problems in continuous speech recognition is due to co- articulation, the way in which speakers tend to slur or glide words into each other.

Price (1969) questioned whether a continuous speech recognisers would ever be possible, at least not until one could give a computer the intelligence and linguistic competence of a native speaker of a particular language. In an infamous letter to the Journal of the Acoustical Society of America, he claimed that the field of speech recognition was dominated by "mad professors and untrustworthy engineers".

The late 1960's and early 1970's saw a rapid growth in the study of artificial intelligence. It was felt that advances in such fields as syntactic parsing, semantic analysis and pragmatic analysis could be usefully applied to the problem of ASR, as could emerging theories of the sound structure of English (see Chomsky and Halle, 1968).

The ARPA Speech Understanding Project

In 1970 a group of scientists, from various disciplines, met under the auspices of the U.S. Defence research projects association to discuss the possiblity of developing intelligent ASR devices. They proposed a five year, $15 million project to develop continuous speech ASR systems of large vocabularies, incorporating some degree of intelligence.

As was said earlier, humans use information from all available sources when they are interpreting speech, i.e. from the speech signal itself, they can get information concerning the speakers mood, age, educational level etc. as well as the actual linguistic information conveyed in the words, or they can get information from the social and environmental context in which the speech is produced. If ASR devices could incorporate some of this extra information in their recognition algorithms, then, it is argued, the accuracy of the recognition will be improved. This improvement might not take the form of high individual word recognition accuracy, but might be seen in high levels of semantic interpretation, i.e. recognition of the meaning of the spoken command.

A number of ASR devices were developed to satisfy these goals. The most successful were the Harpy and Hearsay-II systems. These systems are described in the following section, together with earlier examples. It is not thought useful to include more modern systems, as these are rapidly being superseded. Therefore, the earlier devices will be described to give an idea of various approaches to speech understanding.

Examples of Speech Understanding Devices

i.)Vicens (1969).

Vicens (1969) demonstrated a conti

knowledge of the structure of the language that it used.

ii.) Hearsay-I.

Hearsay-I (Erman and Lesser,1980) was a voice controlled chess player. A

the device achieved a recognition accuracy of 79%.

iii.) Harpy

recognised. A search path technique is employed in recognition of phrases as well. This technique uses a strategy known as beam search. Beam search successively scans a number of alternative steps, which are possible from the search point. By searching alternatives, the need for backtracking is removed. Problems of where to focus the search are constrained by only permittingthe search to proceed from left to right.

In these speech understanding systems (i. to iii.), knowledge is seen as an implicit part of the matching process, i.e. knowledge sources set parameters which constrain the search space. Discussion of the performance of expert spectrogram readers above suggested that such parameters are explicitly modified in the light of new information. In other words, some sort of decision process is used to control the recognition strategy.

iv.) Hearsay II

Hearsay-II (Reddy, 1976) used a 1 011 word vocabulary to allow the control of

it ran in several times real time, and so could not seriously be considered for real applications.

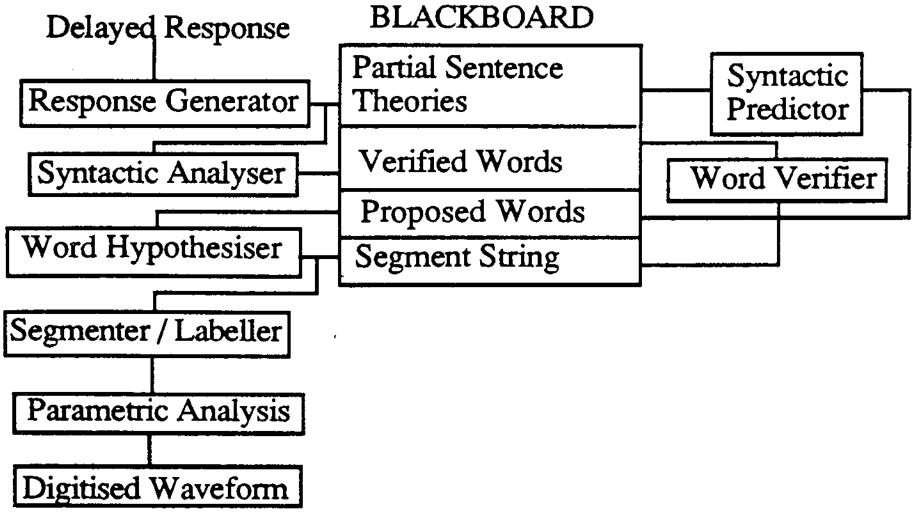


Figure 4.6: Schematic Diagram of Hearsay-II.

However, it was capable of correctly recognising the semantic content of a command 90% of the time. This means that it usually performed the appropriate action, even if it did not correctly recognise all the words spoken. Bearing in the mind the complexities of speech and the problems associated with individual word recognition, it is possible to argue that Hearsay-II, in fact, performed very successfully. If one issues a command to a system, the most important action one wants the system to perform is the carry out the command accurately.

Problems of Speech Understanding

One common criticism of the ARPASUR research programme is that it attempted to force ASR to run before it had mastered crawling! The linguistic knowledge sources are important, but the front end of the system (the acoustic phonetic conversion process) needs to be developed to provide adequate input to the recognition process. Speech under standing research appears to fairly low key at present, with researchers concentrating on improvements to the recognition process, e.g. phoneme based recognition.

In order for ASR devices to reach a level anywhere near human performance, it will be necessary to develop and refine the techniques which utilise knowledge sources in the recognition process. This development can take one of two forms.

The approach of the Hearsay II system utilised several sources of knowledge. As mentioned previously, it is surprising that researchers into ASR did not take a leaf out of artificial intelligence research and aim to emulate the performance of expert language users. Humans incorporate knowledge from a whole arsenal of different sources (Frauenfelder and Tyler, 1987).

If ASR is to improve, it will need to incorporate different knowledge sources into the recognition process. At one level, this means exploration of natural language programming and artificial intelligence techniques to provide 'top down' information to guide the recognition process (see, for instance, Winograd,1980). Assumptions underlying natural language use do, however, present a problem in the design of suitable dialogues between ASR systems and users, as we shall see in chapter ten. This means that, in addition to incorporating linguistic knowledge into the recognition process, it will be necessary to include knowledge of the task domain and the structure of the dialogue.

Knowledge of task domain offers the additional benefit of assisting in the resolution of linguistic ambiguities. This means that the restricted set of language used to perform specific tasks in specific applications will tend to assign very restricted meanings to the words in the vocabulary. Where the words used may carry several meanings in normal usage, in the task domain, they will be less prone to ambiguity. At another level, refinement of ASR, to incorporate more knowledge sources, will entail combining different recognition algorithms to produce more efficient signal processing and 'bottom up' information. For instance, Bowles et al (1988) show that combining dynamic time warping with hidden markov modelling can produce a recognition algorithm which is consistently as efficient as the best performance of the techniques working separately.

Conclusions

Automatic Speech Recognition devices are capable of reasonable performance on vocabularies of between 50 to 1 000 words, and for speakers who have trained the device before using it. It is still a developing technology, with a rapid increase in connected speech systems on the market, and with capabilities for larger vocabularies being developed.

It has been suggested that control room systems require the use of speaker dependent devices, in order to allow vocabularies to be designed and modified according to system demands. This raises the question of how to carry out the enrolment of the device which is discussed in some detail in chapter thirteen, and how to design the vocabulary. Although the technology described in this chapter is very limited, when compared to human speech handling capabilities, there are many applications of speech recognition in avionics, industry and telecommunications, with many more applications being designed and tested (see chapter six). With careful system design which combines an understanding of the basic principles of ASR, with human factors guidelines, it is possible to design ASR systems for control rooms (see chapter nine).

CHAPTER FIVE

ASSESSING THE PERFORMANCE OF ASR DEVICES

Although ASR is very limited when compared to human speech perception, it is able to perform well enough to be of some use in industry. Before using an ASR device, it is necessary to have some idea of how well it does work. This has proved to be quite a difficult prospect, with many researchers unable to define adequate performance standards for ASR technology. This chapter reviews the suggested approaches to the assessment of ASR and proposes a combination of techniques, based around task performance will provide enough information for device selection and system design.

Introduction

Before any decision can be made concerning the viability of ASR in applications, it is necessary to ask questions concerning the performance and capability of ASR and of specific devices, such as:

* How well does the ASR device work?
* How well need it work in order to be effective?
* Can the benefits of ASR be demonstrated in quantitative terms?
* Can laboratory test data predict field performance?

The answers to these questions require data derived from some form of assessment procedure. It is necessary to assess ASR devices in order to provide meaningful measures of their performance, using replicable testing procedures.

Broadly speaking one wants to assess Automatic Speech Recognition devices to see how well they perform in comparison with a manufacturers set of specifications, or in comparison with other ASR devices, or to see how well they will perform in a particular environment on a particular task or set of tasks. Each of these types of assessment obviously require different strategies and units of analysis. Arguments about what would constitute a meaningful, single standard score confuse the types of assessment presented here.

The use of standardised tests and benchmark scores may provide useful data in the comparison of devices, and this would provide useful sales 'propaganda' for manufacturers, but not necessarily give a buyer much information on how well the device will perform in a chosen work environment. Such information can only really be obtained by careful feasibility studies, which would assess tasks and their environment and define suitable performance criteria for selecting an ASR device.

The alternative, and one currently most popular because it is easy to carry out, is the " buy and try " approach. Here an ASR device is bought and then tried in various versions of the target tasks. The problem with this approach is that it does not consider a great many of the variables which may cause problems later in the actual use of the device.

Most manufacturers offer consultation, which although usually skilled in selecting tasks for <u>their</u> ASR device, might not offer the buyer the unbiased selection strategy they hope for.

Basic Performance Measures

Effective performance of an ASR device can be determined by the appropriateness of its response to input signals, i.e. if one says "open", one will expect the device to,

a.) show that it has recognised the word "open" and not some homophone of the target word, and

b.) to send the command to the host computer to perform that action associated with that particular word.

One can define the "ultimate measure of performance" of an ASR system as the effectiveness with which it performs the tasks for which it is used (Spine et al, 1984). But an indication of more fundamental performance can be obtained by taking the first part of this performance parameter; how well a device recognises speech by using a measure of Recognition Accuracy (RA).

Most ASR devices are given a RA measure by their manufacturers. This measure is supposed to indicate how well the device performs, by giving a value of how many words the device can recognise per hundred words spoken to it. Thus, RA is expressed as a percentage.

However, what this figure fails to tell the prospective user is exactly what it purports to state, ie how well the device will perform! What cannot be included in such a figure are details of how well the device will perform under varying environmental conditions, how well it will cope with different speakers, or how well it will cope with different vocabularies.

It is perhaps not all that surprising to discover that manufacturers conduct their tests for RA under as near perfect conditions as possible: the test vocabulary is specially designed to reduce the possibility of confusion between words; experienced speakers are used to test the device; the environment is that of the quiet, clean laboratory... given this, the cynic may be surprised to see that the device can manage a RA of only 98%.

When ASR systems are introduced into the field, it is not uncommon to find a deterioration in RA. This is often due to the fact that users have received little or no training in the use of the device, and that the vocabulary has been poorly designed. The environment in which the device is to be used also contributes to changes in the RA of the device.

This discrepancy between RA promised by the manufacturers and that found in the field often leads to disillusionment. Furthermore, with most commercially available systems offering RA in the region of 97%+, it is difficult to see how systems differ in terms of performance, until they can be tested on other criteria. For example, it is a common observation that a device scoring 80% RA on an 'easy' task need not be worse than one scoring 40% on a 'hard' task.

It has been suggested that rather than look at the RA of the device, one ought to look at how well the device deals with errors. It is possible to define three types of errors in terms of ASR device performance (see chapter twelve). Each of these errors can be expressed as a percentage of the total number of errors a device makes. Substitution errors occur when the input word is misrecognised as another word in the device vocabulary; false rejection errors occur when the device rejects a legal word; and false acceptance errors occur when the device accepts an illegal word or sound.

These error percentages can be used to tell buyers how well a device will perform in respect of error handling. However, the scores do not convey any information about the nature of the errors beyond the crudely defined classes, ie one does not know what words/ sounds are falsely accepted and how these relate to the vocabulary; one does not know whether word boundary violations occur and, if they do, how they will effect the device's performance; nor can these figures reflect the overall distribution of errors in the devices operation.

Thus, although both RA and Error Percentages can provide convenient basic measures of performance, they do not convey enough information for adequate assessment to take place. The concept of RA has been questioned by the majority of researchers in the field of ASR and now seems to be used mainly by the manufacturers for advertising purposes. But it is proving very difficult to develop adequate, standardised measures of ASR performance which will allow direct comparison of different systems.

Confusion Scores

An alternative to using RA scores or error percentages, is to provide some indication of where confusions are liable to occur in the vocabulary. If it is possible to predict that certain words will be mistaken for others on a regular basis, then one could redesign the vocabulary to reduce this effect. Waterworth (1984) proposes the use of recognition trees to assist in this decision process. Here speakers are asked to repeat the words in the vocabulary a number of times, and the total percentage of confusions between words is indicated. Figure 5.1 illustrates a confusion tree of the digits 1-10.

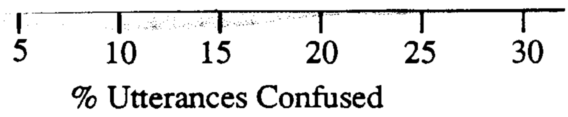


Figure 5.1: Confusion Tree for the Digits 1-10

[From Waterworth,1984]

Although the confusion tree shows where confusions occur in the vocabulary, it cannot show that such confusions are often asymmetrical. That is, a confusion tree cannot tell us that words are confused in certain directions rather than others, e.g. to say 'five' is confused with 'nine' does not indicate whether the recogniser will return 'five' for 'nine' more times than it will return 'nine' for 'five'. This can be an important factor for some vocabularies, with key words being confused more often with other words. This pattern of confusion can be displayed using an alternative means of confusion presentation, known as a confusion matrix (see figure 5.2).

As part of the research for this thesis, an assessment of a new device developed by Marconi was performed. This device was the Macrospeak. A full report of the assessment and findings can be found in Usher (1988). One of the tests employed investigated the likelihood of confusing the names of mills in power station displays. At present mills, for grinding coal to 'pulverised fuel', are identified by the letters "a" to "e". A confusion matrix, shown below, was developed to indicate where possible confusions would occur when displays of each mill were called onto the screen.

Although the confusion matrix can tell the user what types of substitution error are likely to occur, it does not provide any information about how well words are being recognised. In chapter one, the concept of pattern matching was discussed, and the principle of distance scores defined. When a spoken word matches a template, it is highly unlikely that they will be exactly the same. Rather, there will be some difference between the two words. This difference is expressed as the distance between the spoken and stored words. The lower the distance score, the better the match.

		Word Spoken							
		Mill A	Mill B	Mill C	Mill D	Mill E	Mill G	Mill V	Mill P
Word Recognised	Mill A	7	▨	▨	▨	2	▨	1	▨
	Mill B	▨	5	▨	▨	▨	▨	5	▨
	Mill C	▨	1	8	▨	▨	▨	▨	▨
	Mill D	▨	▨	▨	4	▨	1	▨	5
	Mill E	1	▨	▨	▨	8	1	▨	▨
	Mill G	▨	1	▨	▨	▨	7	▨	2
	Mill V	▨	4	▨	▨	▨	▨	5	1
	Mill P	▨	1	1	▨	▨	▨	▨	8

Figure 5.2: A Confusion Matrix showing the Number of Occasions Spoken Words were Recognised Correctly or Misrecognised with Other Words.

It is common for manufacturers to quote mean distance scores as performance measures, but these should be treated with caution. It is not possible to investigate the process by which a template and a recognised word have been matched. We can only hypothesise how well they will match. Distance scores offer little information beyond such hypotheses. By quoting a figure, it appears that one is receiving a definite measure of performance, but unless one can adequately interpret the figure, it is meaningless.

Methods of Measurement

RA, Error Percentages, Confusion Trees, and Confusion Matrices are all means of indicating the performance of ASR device, but they give no indication of how to measure the overall performance of the ASR system. This section reports some of the more common assessment techniques, but it is important to note that standard techniques are still in the process of development. As McCauley (1984) points out, the development of a standard for comparing recognizers is difficult because no single set of performance criteria will be appropriate for the broad range of potential applications of ASR.

One method which could offer much potential was developed by Lea and Woodward (1984). This is known as Relative Information Loss (R.I.L.), and relies on the collection of several types of error. By weighting the various types of error in terms of criticality to the task, it was possible to produce a single metric based on information theory measures. This would extend the basic measures based on error rates to allow predictions of performance to be drawn.

When one assesses the performance of a given ASR device, it would be sensible to begin with an assessment of the variables which are most likely to be the cause of variation in recogniser performance. These can be grouped under the following three factors:

a.) Environmental Conditions: including the effects of noise, the type of microphone used and it's placement, interfacing between the ASR device and other equipment, etc.

b.) Speaker Variability: including expected variations in the potential user population in terms of age, sex, regional accent etc.

c.) Vocabulary Characteristics: including the type of vocabulary likely to be used and the possibility of confusion between vocabulary items, and the type of dialogue to be used.

Other factors, which are of equal importance in the assessment of ASR devices, are the type and frequency of errors which the device makes, and such temporal factors as the overall device response time.

Given that all these factors contribute to the performance of an ASR device, it would seem very difficult to begin assessing the systems performance: how can one control for so many factors?

Spine et al (1984) combined a number of factors which they proposed would contribute to the performance of an ASR device. The overall performance was calculated over the various combinations of these factors, and regression analysis was used to tease out the factors which effected, and interacted with, each other. The factors which Spine et al (1984) uncovered were:

i). Number of passes per item at enrolment;

ii). Vocabulary size;

iii). Level of rejection threshold;

iv). Difference score from matching process.

These factors were combined with two types of vocabulary. One used a "worse case" set of highly confusable words, e.g. 'race' vs. 'raze'. The other used a "best case" vocabulary of the ICAO alphabet, digits, and simple command words.

These factors, although relevant to the studies of Spine et al (1984) are questionable. We have already pointed out that the concept of "difference scores" is of dubious utility, and suggest that it does not provide a viable means of measuring performance.

The issue of vocabulary size is considered in chapter six. While specific applications will require different size vocabularies, a device offering several thousand words of vocabulary need not be better than one with a few hundred words of vocabulary. Few industrial applications require large vocabularies (see chapter six). Researchers at Hewlett Packard have shown that performance deteriorates with increasing vocabulary size in accordance with:

$$P= v-1/ v$$

[v = vocabulary size]

This predicts quite small changes in performance over increasing vocabulary sizes. One would expect that the probability of errors occurring to be proportional to the vocabulary size. This is true of human speech perception (Miller et al. 1951). Doddington (1980) played twenty subjects a number of isolated words of human speech. The results showed that an increase from fifty to fifteen hundred words only produced an increase of 0.4%. When the vocabulary was increased to twenty six thousand words, the error rate increased by 1.3%.

One would expect the unit's performance to change in relation to the size of the vocabulary in use. A simple, although time consuming, test would compare the performance of the device using varying sizes of vocabulary. Meisel (1986) discusses a study in which vocabularies of two and five thousand words are compared. The results show that the larger vocabulary does not necessarily yield better performance (see "Office Applications" in chapter six).

Vocabulary size can be defined in terms of the number of words in the vocabulary, or in terms of length of individual words. It is often assumed that there will be better performance for multi- syllable words rather than monosyllable ones, but this effect must surely be due to the 'phonetic richness' of the words used (how distinct the patterns of sounds are in the word). One would suspect that there will be an optimum number of syllables above which performance may deteriorate; there are limitations on the number of words a user can produce in one pass.

Effect of word size can also be examined using variations in the length of frames used for segmenting the speech input. Finally one needs to look at the optimum phrase length which the device can handle in order to design an appropriate vocabulary. Some words, e.g. those which contain stop consonants, may prove to give the device difficulties. Can it handle words which have a short time gap between phonemes, e.g. repea - t, without treating them as two words?

Number of passes per enrolment can be assumed to effect performance. This is investigated in chapter thirteen. While one could use this as a measure for comparison between devices, it will not provide a performance measure so much as a 'setting up' measure. One could enrol the device a number of times, recording the RA after each set of enrolments, until one or other of the devices reaches a desired recognition level. This seems to be a time consuming process and ultimately depends on the use of recognition scores, which have already been dismissed as performance measures.

Some ASR devices allow users to modify the threshold at which words are recognised. This means that a specified distance score must be exceeded before a word is accepted. There is obviously a trade off between number of errors incurred and level of threshold; the higher the threshold the more errors will occur. One could compare devices in terms of performance at different levels of threshold, but unless they use similar recognition and matching algorithms, alteration of thresholds on different devices will inevitably result in unquantifiably different alterations in performance.

Approaches to Assessment

The majority of researchers favour one of the following three possible approaches:

i.) Assessment using prerecorded databases of speech.

ii.) Comparing the performance of the ASR device with that of human listeners in controlled environments.

iii.) Assessment of the device in the field.

However, all three approaches rest on the fallacious view that the performance of an ASR device can be quantified into a single figure.

i.) Assessment using Prerecorded Databases

If we consider the factors which depend the recognition algorithms ASR devices might use, one can imagine that certain databases can provide test vocabularies to assess performance of several ASR devices.

The database could be constructed using a general vocabulary which is suitable for all applications. This would contain digits, and letters of the alphabet spoken by a range of different speakers. Many applications will depend on a selection of key command words. This could be catered for a database which uses task specific vocabulary. However, neither the general nor task specific vocabularies can provide clear scores of overall performance across different recognisers in different conditions.

Databases can be constructed which use principles from diagnostic rhyme tests (House et al. 1965). This will provide vocabularies based on sets of minimally confusable pairs of words, e.g. "pa" vs. "ba", which can be used to indicate which words or segments of words are either missing in the analysis, or else overemphasised. But, as Johnston (1986) points out, this will only indicate where the deficiencies are occurring, not offer a quantification of them.

Using a database could appear beneficial at first glance, but it has some drawbacks which might lead one to question whether the time it takes to construct such a database could perhaps be better employed doing something else. If one wishes to use a database to compare several different units, then one is confronted with the problem of what would constitute an adequate database for each of the units to be tested. If one wishes to assess the performance of a specific device, then one is faced with the problem of what would be an adequate database to represent all the variation that the device is likely to encounter in the prospective user population.

This would tend to suggest that the database needs to be very large indeed to cope with all the factors that can effect the performance of the device. These factors include,

* microphone type and placement;

* speakers age, sex, accent, speech rate;

* vocabulary size and complexity;

* phonological variation in the vocabulary and in the speakers' style of speech;

* environmental factors;

* electrical conditions, such as interfacing between ASR device and other systems.

Peckham and Knight (1984) propose that small databases be constructed which contain examples of speech under certain conditions. This will allow assessors to select only examples of test material which will be useful for the application domain in question.

In addition to the problem of how large the database need be, one is faced with the complicated problem of how to use the database to effect changes in the recogniser's performance. If the device fails to meet the specifications defined by the database, and if one is able to change the performance of the device to come up to these specifications, one is still not able to say how well the device will perform in the field or with respect to factors not included in the database. Finally, the changes which one makes may well only increase the recogniser's ability to deal with speech in the database, rather than speech in general.

ii.) Comparing the Performance of the ASR Device with that of Human Listeners, in Controlled Environments

Some researchers suggest that the performance of an ASR device can be assessed quite adequately by comparing it to the performance of human listeners in specially controlled conditions.

Recordings of acoustically similar pairs of words, e.g. 'deed'/ ' bead', masked by white or pink noise, are played to human listeners (Ohala, 1982). The listeners are then asked to rate how similar or dissimilar the pairs sound. Then these figures could be compared to the confusion scores given by the ASR device. This comparison could then yield a quantifiable measure of the unit's "phonetic resolving power".

Moore (1977) proposed that a model of human speech recognition, under varying conditions of noise, could be used to mirror the performance of an ASR device. This is computed into a Human Equivalent Noise Ratio (HENR), and in addition to offering a means of measurement, also provides a standard level of performance. Miller and Nicely (1955) demonstrate that human recognition of speech tends to follow certain patterns of confusion in relation to different levels of noise.

This seems to be quite an effective way of assessing how well the device will perform in the discrimination of acoustically similar words, especially as the performance of human listeners can be used to provide a benchmark. However, the technique cannot deal with any of the other factors which the device will come up against. ASR systems do not process speech in the same way as humans do. Whereas ASR systems tend to concentrate entirely on the acoustic waveform for the information that they use, humans use as much information as they can and process speech using a combination of bottom- up (acoustic driven) and top- down (hypothesis driven) processing strategies (Marslen- Wilson, 1987). One also needs to bear in mind that the performance of human listeners may not be constant and so scores of several listeners should be averaged to provide a benchmark score.

iii.) Assessment of the Device in the Field

Thomas (1987) points out that the majority of current assessment techniques are content to use the ASR device in it's 'test mode' rather than assessing how well it will perform as part of a device designed to perform a specific task in a specific environment.

One could imagine that assessment could involve comparison of rate of data entry using the ASR device with other devices. The problems here are that because the data will probably be encoded in different ways by the user and the device, meaningful comparison is often difficult to achieve, e.g. if one compared a task of entering words using a keyboard and an ASR device, then one would need to consider whether to examine the results in terms of number of actions (single key strokes for each letter vs. utterance of each word), or in terms of time taken to complete the task (which would vary as a function of how many actions the user needed to perform top complete the task; speak one word or type five letters). This issue is investigated more fully in chapter seven.

Peckham (1986) suggests that one assesses the use of an ASR device in terms of the "transaction time" necessary to perform an operation. Transaction time will be " measured from the moment the user receives the stimulus to carry out the task to the time he decides that it has been carried out satisfactorily." The problem with this is that the time taken for the user to make decisions concerning the task he needs to carry out etc. might effect the "transaction time" measure. Visick et al (1984) have, however, shown that the time it takes for an ASR device to recognise speech can be an important factor in the overall performance of a parcel sorting task.

Another method of assessing "system performance", as opposed to device capability, is to use a measure based on 'secondary task' performance. This methodology, and the theory behind it, is described in more detail in chapter eleven. Briefly, it is based on the assumption that as the amount of work a person is called to perform increases, their overall performance will decrease. It would be possible to assess whether ASR can be used to maintain an acceptable level of performance, in comparison with other input media, in high workload situations.

Conclusions

During the writing of this thesis, I assisted.in the assessment of an ASR device marketed by Marconi. This assessment took three speech styles to measure the devices performance: speaking a phrase at different speech rates, to see how 'continuous' the recognition performance was; speaking words from a highly confusable vocabulary; issuing commands to a voice controlled game. Full details of the results are reported in Usher (1988). One major problem we found concerned the interpretation of the results we had obtained. What did the performance scores we had measured tell us? There are, at present, no agreed standards of ASR performance, and so assessment must either take place in the context of device comparisons, or rely on the intuitions of the assessors. Speech databases will provide some consistency of assessment measures, but they are subject to problems discussed above.

One must agree with Thomas (1987) in his conclusion that,

> " Because speech recognition is a complex action, including not just the recogniser itself, but also the interaction between the speaker's environment, the speaker's own voice characteristics, the higher level processing undertaken by the driving computer, and the type of feedback given to the speaker, it would be unreasonable to expect that a single performance figure could be produced which would be taken to be the recogniser's performance."

Cotton (1981) has suggested that rather than assessing the performance of an ASR device with respect to standardised measures, one should look at ways of classification in terms of the number of speakers the device can handle, the vocabulary size, noise tolerance, etc. This would allow some sort of simple performance comparison to be carried out, and is usually the accepted method of presenting performance data when devices are compared in articles.

There are a number of potential applications which one can define in terms of whether they will require isolated or continuous word recognition, what sort of noise is likely to be present, how many operators will be working at once etc. These factors could be used to suggest permutations of performance parameters to give some indication of performance under varying conditions.

Using the " phonetic discrimination test" proposed by Ohala (see above) would provide a benchmark of sorts, by using human speech recognition performance. This test could be augmented and improved by tempering the results with actual performance figures from a range of systems.

The microphone does a great deal of work for the unit, one could ask how will the device perform if the microphone is changed to a low quality one or if the microphone is not positioned correctly ? Although the microphone chosen for actual use will be one which will optimise device performance, it could be argued that the better the microphone the easier the task of the ASR device to discriminate speech from noise. If one tested the device using a poor quality microphone one could see how noise intrusions effect performance.

One needs to begin with a consideration of why one is assessing the device, i.e. to compare it with other devices, or in terms of manufacturers specifications, or in terms of performing specific tasks. In general, one will want to assess ASR devices on the latter two dimensions, and the approaches and ideas presented here are geared towards such studies. The former dimension is most useful for manufacturers to assess their competition. However, for buyers, comparison of devices needs to come at the end of a feasibility study, which would provide the definition of criteria for adequate performance of an ASR device in the prospective environment. Manufacturers specifications would provide information for initial selection, and then devices could be assessed in terms of the other two dimensions.

What the approaches reported above do not seem to consider in any detail is the effect of using the ASR device as part of a Human Computer Interaction (see chapter one). Writers seem to agree that the main advantage of using ASR devices does not lie in their ability to act as keyboard emulators, but in their ability to recognise speech, and allow the exploitation of a characteristically human mode of communication. Thus, rather than assessing the performance of the device alone, one needs to consider it as part of a communications loop. This would require that the type of feedback and prompting provided by the device be assessed to ensure that adequate user support is given.

This last detail is usually left to the buyer to set up, and one can add to this the factors of training the users, device enrolment, vocabulary and dialogue design, and feedback considerations. It is these factors which will make the biggest difference to the performance of the ASR device, and, obviously, these are the factors which will most vary across different tasks and companies. ASR device assessment needs to be carried out as part of an ongoing research project, which is able to assess the selection and implementation of the device in a specific environment for at least two years. This will provide information on device performance which will be far more detailed than a single performance measure can indicate.

CHAPTER SIX

CURRENT APPLICATIONS OF ASR

Even though new technology is being developed, the basic principles of ASR described in chapter four seem relatively constant. We will now turn to the issue of ASR application, and consider examples of how ASR has been used in industry, avionics, tele communications, office environments, and for disabled people to illustrate its potential benefits.
The applications from industry are assumed to bear greater similarity to those found in control room systems than applications in other areas , and consequently are considered in greatest depth.
From the discussion, points concerning the selection of appropriate applications of ASR in control room systems are raised.

Introduction

We have seen, in chapter four, that commercially available ASR devices are capable of recognising human speech under quite tight constraints. The majority of commercially available ASR devices are still of the isolated word recognition type (although there is a very rapid increase in the availabilty of connected speech devices), which requires pauses to be introduced by the speaker between each word or phrase. The vocabulary range of such devices is limited to between fifty to a thousand words. The type of ASR device most suitable for control room operations will be user dependent, as these allow the command vocabulary to be tailored to specific applications and situations. However, these devices require the process of enrolment. Enrolment requires around three pronunciations of each item in the vocabulary, in order to provide suitable reference templates (see chapter thirteen). This could be very time consuming and irritating for new users.

It is the aim of this thesis to demonstrate that these criticisms are not, in fact, problematic for the introduction of ASR in control room systems. The use of ASR tends to conjure up science fiction images in prospective users (see chapter one), which are not only not warranted by the technological capabilities of the devices, but are also based on mistaken understandings of the requirements of the tasks in question.

In its most basic form, ASR could be described as providing "voice buttons" for computer input and operation. This would allow keyboards to be replaced by ASR without drastically effecting system operation. ASR would simply emulate the previous keyboard, to allow the operator to input strings of digits or short command phrases. As we shall see below, there are very many industrial applications which benefit from such use of ASR. The use of ASR removes the need to type in data, and so frees the operators hands for other activities, such as baggage handling. Such applications only require isolated word recognition.

Data entry of digits is very simple and easy to perform using ASR, but there are also applications which require command strings to be spoken. One might believe that the limited vocabulary of current ASR devices would not be able to cope with such demands. However, research into the potential application of ASR in fighter aircraft, reported in Hill (1980), found that thirteen thousand distinct control sequences could be performed with a vocabulary of just fifty four words. Therefore, with judicious vocabulary selection and design, it is feasible to use ASR for quite complex activities.

The introduction of ASR in any working environment will require consideration of several factors. Chapter five showed how assessment techniques should consider the majority of the factors before the device is used in the work environment. However, after assessment, it is important to be aware of the problems of background noise. Background noise could either originate in the work environment, such as the sounding of alarms, or in the operator, such as coughing etc. or 'babble'. This later term refers to the practice of talking over a live microphone to other operators, or to oneself. Whichever form 'babble' takes, it will disrupt ASR performance.

Environmental background noise can be dealt with by a number of factors. As the applications considered demonstrate, the most common use noise cancelling techniques. Operator background noise requires consideration of training and of microphone operation, with the requirement of a means to turn the microphone off when the ASR device is not in use.

The following sections discuss various applications of ASR in industry, avionics, offices, telecommunications, and for the disabled. Of prime interest to control room design is the use of ASR in industry, as there seems to be the closest link between these two fields. Consequently, this is the area which receives the widest coverage.

Industrial Applications of ASR

Owens Illinois Corporation has been using ASR systems, in the task of inspecting faceplate panels of colour television sets, since early 1973 (Reddy, 1976). In this system an inspector first types the following information into the computer: the type of t.v. to be examined, employee number, shift, date, time. This data is then used as the heading to a summary report, produced at the end of the inspection task.

The inspection task requires a sequence of measurements to be taken, which the inspector is prompted through by commands on a visual display. The display gives the name of the measurement to be taken, a nominal value, tolerance limits, and the data being entered.

In order to measure all the components, the inspector has to manipulate the faceplate and use hand held meters. If he was required to record each value by hand and then calculate the out of tolerance values for each component, the task would be very time consuming and laborious. Using ASR leaves the inspector's hands free to use the measuring devices and to manipulate the faceplate.

The system uses a simple vocabulary of digits and a few command words, e.g. the inspector is required to say the difference between the measured value and the nominal value: if the measured value is 20.041 and the nominal value is 20.000, the inspector need only say "41". He moves on to the next measurement by saying "go". This allows the inspectors to work at their own pace.

If the measurement is out of tolerance, the display will show <out of max/min>. At this, the inspector can either say "erase", to erase the last spoken value and input a new one (this provides a method of checking that the system has not made a mistake in recognition); or he can say "over ride", which will ensure that the part number and its value are recorded as faulty in the final report.

At the end of the inspection run, the computer prints out a summary inspection report with all the information from the run, along with data for mean, and range of, faults as a check for batch quality. Thus, the inspector is saved from performing various tasks, such as recording measurements and completing a final report by hand. The use of ASR allows the inspector direct access to the computer, thus relieving them of the task of calculating tolerance values or remembering digit strings for long enough to them record, while performing several measurements.

A similar system is in use in a factory (see Martin, 1976) to check the quality of the ring pulls for tin cans. A difference between the two systems is that whilst the t.v. inspection procedure is performed by one person at a work station with all the necessary tools to hand, the ring pull inspection task takes place on a conveyer belt; several people work on the same tasks. This means that inspectors may have to wait for equipment, and structure their inspection procedure to coincide with the availability of tools. The system allows some flexibility in the ordering of the tasks to be performed and the data input.

Both of these inspection task applications show that ASR can be used to reduce an inspector's physical and cognitive workload. Data can be spoken directly into the computer, allowing them to be recorded while the inspectors' hands are busy and while their eyes are fixed on a specific task. Flanagan (1976) suggests that in situations such as checking complicated wiring installations, an inspector can use a two way radio link to a computer to enter measured data and be guided through the inspection process.

These applications also illustrate how ASR can increase the throughput of inspected components. Poulton (1983) tells of a bottle making plant which uses voice input to its numerically controlled machinery. This saves time spent in the preparation of tape input: one company with 18CNC machines cut programming time by 65% to 90% (depending on the complexity of the part being machined).

Backer (1984) reports the use of a connected speech recogniser in a baggage sorting application, which has been in operation since 1973. Operators in a baggage sorting task read the destination off the luggage tag into the system microphone. The system then conveys the baggage to the correct destination gate. When this process was performed manually, an average of 12 bags per minute per operator were sorted; with ASR the average per operator increased to 30 bags per minute. There was also an increase in the accuracy of sending bags to the correct destination gate which amounted to a performance increase of around 33%.

Visick et al. (1984) report the use of an ASR system in a parcel sorting task. They found that, in a task involving the entry of destination names, performance was better using manual controls than using ASR. In a second study, they found that if subjects were required to sort parcels in addition to entering the destination, ASR produced faster performance than manual data entry. However, Visick et al (1984) reported an error rate of around 40% for ASR compared with 1% for manual data entry (in both studies). These findings can be explained by the fact that in their study, Visick et al (1984) asked the operators to use the names of the destinations of the parcels as input to the system. This led to some confusion between similar sounding names. In a more recent study, Frankish and Jones (1987) used the ICAO alphabet in a parcel sorting task, and found the error rate was much lower and that the task was performed faster than in the Visick et al (1984) study. This illustrates the need to design a vocabulary of minimal confusability and that the characteristics a particular system must be taken into account when planning an implementation.

Lerman (1980) points out that in the assembly of high reliability electronic equipment, there is a need to trace the history of every component back to it's first stage of manufacture. A system which uses 15 voice data entry points connected to a minicomputer is described. At each stage of manufacture data concerning each component, are input by voice. These data are then analysed by the computer. When the component is complete, the data are passed on to a mainframe computer. The system operates on a vocabulary of 47 isolated utterances (mainly digits, the phonetic alphabet and a few command words).

Thus, ASR can be seen to require fairly small vocabularies. Most industrial applications can use vocabularies well within the range of fifty to a thousand words for commercially available ASR systems. Nye (1980) believes that most industrial applications require around 75 words in operational vocabularies. In a survey of contemporary applications, he found the distribution to be as follows: 8% of applications required 12 word vocabularies; 88% required 75 words; and 4% required 76+ words. 66% of connected word applications required around 200 words.

Application	Improvement over manual	Vocab. Size	Savings	Source
Baggage handling	33% over manual	ICAO+ digits	> in speed+ accuracy	Backer (1984)
T.v. faceplate inspection	Hands/eyes free	Digits + 5 commands	> speed	Martin (1976)
Data capture in elec. man.	Direct data capture	47 items	____	Lerman (1980)
On line inspection of cars	Hands free DDC	34 phrases	" Improved Performance"	Byford (1987)
Goods Inspection in Warehouses	Hands/eyes free, Mobility.	Goods' codes.	___	Rehsöft (1984)

Figure 6.1: Table of ASR Applications in Industry

Lea (1980) questions whether most applications actually need connected recognition capability. 80% of all industrial applications use isolated word recognition (Reddy, 1980). Furthermore, with the increase in the level of technology in isolated word recognizers allowing the recognition of quite long phrases, i.e. of 6 or 7 words, it is difficult to know where to draw the line between the two types of system (a problem further complicated by the reduced pause times required between isolated words on some systems).

Rehsöft (1984) describes the application of ASR to an online shipping system at a car component distribution centre. By using a lightweight radio headset, operators at various work stations are able to enter codes and weights directly to the computer and have some degree of mobility to move packages.

It is important to note that the use of a radio microphone is subject to some restrictions. Peckham (1989) reports the use of a radio link for a car inspection plant. The main restriction he noted was the fact that the Department of Trade and Industry classifies a permanently open, two way radio link as a radio station, which requires special licensing. It would be possible to use the radio link intermittently, but Peckham (1989) questions the utility of a two way link in terms of expense and weight. A one way link would be preferable, from operator to computer, with prompts and feedback appearing on a visual display. Operators interviewed during the field trails reported in chapter eight, suggested the use of LED displays, such as are found in public displays in shopping centres.

An alternative to using radio links is infrared transmission. This requires the placing of several sensors around the workplace to receive the signals. This could be a feasible solution to control room systems, but consideration needs to given to the viability of infrared transmission in terms of the reduction in available bandwidth when compared to radio links.

Cassford (1988), Byford (1987), and Anderson and Gill (1987) report the implementation of ASR systems on the production line at automobile factories. Inspectors are required to check the components used in the vehicle. The operator begins each shift by logging on to a computer terminal keyboard. Next the operator speaks a simple test phrase to calibrate the system to current voice levels. After this the operator speaks his name; this allows validation of ID and ensures that the correct voice templates have been loaded.

The operator then speaks the serial number of the vehicle to be checked; this enables the computer to select the set of checks which are suitable for that vehicle type. The operator is guided through the inspection procedure by a series of synthesized voice commands over a headset (the complexity of these commands can be varied according to the operators level of experience with the inspection task).

ASR in Telecommunications and Consumer Products

When writers consider the use of ASR in consumer products, their imaginations tend to run riot; everything from voice activated microwave ovens to speaking wristwatches have been suggested. Baker (1981) states that,

> "As a status symbol, [ASR] competes with flashy cars
> and designer clothing. Controlling computers promotes
> a sense of power; automatic speech recognition promises
> to make it easy".

While consumers may be swayed by the logic of such homespun rhetoric, industry is rightfully wary. Doddington and Schalk (1981) suggest that the application of ASR to such things as speech controlled t.v. sets,

> "contributes little or nothing to tasks that already
> may be manually performed quickly, reliably and
> inexpensively".

The Alvey report on advanced information technology saw voice as one of the important tools in the ' communications revolution', by making human computer interaction easier for the non specialist. It should be pointed out that for such uses as automatic telephone transaction systems, there still a great deal of work to be done to enable the systems to cope with the wide variety of speaking styles that it will encounter. However, several systems which have been developed look promising (Bruce, 1987 and Peckham, 1986 report research on the VODIS transaction system for train timetable enquiries; Levinson and Liberman, 1981 report research on transaction systems for airline reservations.)

There are some ASR applications in the field of telecommunications which already commercially available and useful. Viglione (1986) reports work on a voice activated telephone, which sells for between $200 to $300. The user 'dials' by speaking the required number, or by speaking the name of the person they wish to call (providing that the name is stored in the system's memory associated with the number). The IVP voice telephone uses a 17 word vocabulary (digits 0 9 and command words).

An even simpler use, and one which does not strictly fall into the category of ASR is ' voice messaging'. Here a message is stored and played back when the user requires to hear it. The system works like a cross between an answerphone and an 'electronic mail' system (see below).

ASR in the Office

The main environment for telecommunication products would be the office. Noyes and Frankish (1989) estimate that between 15 to 20% of office work is involved in spoken communication. McMahan (1984) proposes that the use of voice messaging could save workers time and allow them to structure their work more profitably. Rather than being interrupted by telephone calls, one could use a voice messaging system to record the calls. This is not dissimilar to conventional answerphones. However, voice messaging allows the user to file and annotate stored messages using spoken commands, which could prove highly valuable in busy offices and allow a transcription of telephone calls related to different topics to be collated.

For many workers, typing is a laborious and error prone business. Clearly a workable alternative would be welcome. One can characterise ways in which ASR might replace keyboards on a scale of task complexity. At the lowest level, ASR might be used to annotate documents already stored on the computer. van Nes (1987) found that spoken annotation was far faster and more verbose when using ASR than typing, although there was no noticeable difference in message content between the two modes.

Following annotation, would come information retrieval and data entry. Consideration has already been given to information retrieval from databases over telecommunication networks. Data entry has been shown to be viable in industrial applications above. Starr et al (1988) report the use of ASR in recording information when digitising hydrographic charts. ASR is used in conjunction with a hand operated cursor. The issues of using ASR with other media is considered in chapter seven.

Finally, the most complex application of ASR is the "talkwriter". The first such system was demonstrated in 1956, and could recognise the items "I can see you type this now" (Olson and Belar, 1956). The "talkwriter" represents the ultimate ASR system, in that it is intended to capture human speech and transcribe it accurately. Needless to say, there are many research projects investigating this issue, but it is too early to assess either its success or potential, although Kurzweil is marketing such a device.

There some human factors considerations of 'talkwriters' which are of interest to this thesis. Bahl et al (1983) compared recognition devices with two and five thousand word vocabularies on a three thousand word test vocabulary. Obviously the two thousand word vocabulary could not deal with all of the words, whereas the five thousand word could.

However, the words in the two thousand word vocabulary were the most common ones found in the system application; the remaining thousand words were used very rarely, and could conceivably be spelled by the user if they were not recognised. The five thousand word vocabulary showed a larger rate of errors (chapter five demonstrated the probable increase of errors with an increase in vocabulary). This suggests that there is no advantage in using a very large vocabulary; what is more important is to include the most appropriate words in the vocabulary that is used.

The "talkwriter" could be considered to require connected speech in order to be of any use. However, Gould (1978) has shown that when experts dictate letters etc. they tend to pause and stop constantly. This suggests that a low speech rate is not actually a hindrance, but could support normal composition speeds. Indeed, Carter et al. (1988) found that composition rates using a simulated 'talkwriter' only averaged twelve words per minute. Further, the ASR process need not even run in real time, as experts are used to dictating a passage and receiving the draft later in the day to check (Meisel, 1986).

ASR could be used in other forms of computer interaction. For instance, Schmandt et al (1990) report a voice controlled window system for pcs. In this system, the problem of overlapping windows can be reduced by speaking the name of a particular window to call it onto the screen. This was shown to be a more efficient means of window management than conventional devices, in situations involving several windows. Schmandt et al (1990) argued that this resulted from the fact that their application exploited 'eyes free' use of ASR. However, it is obvious that their results could be equally explained by the fact that ASR does not require specific areas of screen or keyboard to be displayed for use. In this respect it has an advantage of conventional input devices in that it can be used without the user needing to address a permanent display.

ASR for Disabled People

ASR can be of significant benefit for disabled people (see Damper, 1984; Kraat, 1985; or Noyes et al, 1989 for discussions of the range of applications of speech technology for disabled people). An ASR device can be used as a front end to an intelligent speech trainer (Damper, 1984). Users will be prompted to speak a word, and the matching process of the ASR device will compare this with 'acceptable' pronunciations enrolled previously.

ASR can be used to control domestic appliances. Haigh and Clark (1988) described a Voice Activated Domestic Appliance System (VADAS), based on a speaker dependent, isolated word recogniser. VADAS had a vocabulary of seven command words which could be extended to control up to sixteen appliances. Initial trials were disappointing, due mainly to the effects of noise. It was suggested that the use of a desk mounted microphone affected performance, with a great deal of variation in speaker position. This could be solved with the use of a head mounted microphone, but it is difficult to imagine users being happy having to wear the microphone permanently for intermittent control actions.

Lees et al. (1988) report the development of a speech controlled robot arm to assist disabled people in basic eating and hygiene tasks. The main problem that this device is currently faced with is the definition of spatial commands couched in vague terms. The problem of using ASR for identifying and changing the spatial locations of objects could be useful in control rooms. For instance, the status of objects on a wall mounted mimic diagram could be changed using spoken commands. One can imagine a dynamic mimic with objects which could be moved by voice. The main problem with spatial instructions lies in their inherent ambiguity: what does the command "move it left a bit" mean, left of what and how far is a bit?

Sondheimer (1976) has suggested that spatial commands could be used for natural language interaction with computers, providing that the syntactic and semantic structure of the commands was very tightly controlled. This could lead to limitations in the possible range of uses or to problems of unnatural command structures. However, Schmadt (1986) reports the use of ASR for speech control of a dynamic, computer generated map. Users could issue speech commands to move ships around the map, corresponding to the ships position and movement. Problems with commands are tackled using a sophisticated form of dialogue control.

One problem which is often encountered in the application of ASR for disabled people is the effect of changes in the speakers voice due to stress or fatigue (Haigh and Clark, 1988). This is also a problem for industrial applications. As Newell (1986) states, the application of ASR for disabled people tends to present an exaggeration of applications problems encountered by researchers into applications for able bodied people. The use of speech to control a 'talkwriter' could prove very useful to people suffering from spinal cord injuries (Stephens et al. 1988), and Fisher (1986) proposes the use of voice controlled telephone dialling for disabled people. Both examples illustrate that what might be gimmick products for the able bodied can be very useful for the disabled.

ASR in Avionic and Military Applications

A great deal of research has been carried out by the military, particularly in the USA. The majority of research in the area of avionics tends to consider both ASR and speech synthesis. There are currently operational systems which use speech synthesis in alerting and warning pilots of control readings (Aretz, 1983), but ASR is proving harder to implement.

Cotton et al (1983) and Reed (1985) interviewed pilots of fighter aircraft as to what would be the most appropriate applications of ASR in the cockpit. They found that data entry and information retrieval tasks were the most commonly cited, followed by selecting and changing radio frequencies.

In a fighter jet, the pilot is expected to understand and respond quickly to various information channels. He must maintain a high level of attention both inside and outside the cockpit. The pilot acts as a'manger' of many automated systems which require monitoring and control, as well as flying the aircraft (Reed, 1985). ASR is intuitively attractive in the cockpit because it allows the pilot to keep both hands on the controls and to attend to things outside the cockpit.

Reising and Curry (1987) compared ASR with a touch screen using a multilayered menu in a flight simulator for the selection of radio frequency bands. They found that ASR yielded fewer errors compared to touch screen control, and that when using ASR pilots took less time to complete the task than manually. However, when the menu for the touch screen was redesigned to display common functions on the top level, the time consuming task of searching through levels was eliminated from the touch screen condition. This resulted in similar performance times for touch screen and ASR, with ASR producing more errors. Thus, before comparing ASR with other input media, one should optimise performance of all media in order to achieve a fair comparison.

The main argument for using ASR in the cockpit is that, as flying consists primarily of visual spatial and manual activities, the pilots speech channel will be clear. Linde and Shively (1988) investigated the existing use of speech communication, via radio links, in a police helicopter. They found that the pilot and crew in constant spoken contact with each other, with control towers and dispatchers, and with ground units. Indeed, they estimated that between 30 to 50% of flying time was spent in short spoken exchanges, on average lasting twenty five seconds. This suggests that the speech channel is not always clear in the cockpit, and that the introduction of ASR could impinge on a number of existing speech communication tasks. One needs to investigate the proposed domain with care before attempting to use ASR, rather than rely on the assumption the operators inevitably have spare capacity in the speech channel.

The environment of the fighter aircraft is foreboding to ASR use for a number of reasons. There is a high level of background noise. This is known to effect recognition performance, although it is possible to design recognition algorithms which can compensate for the effects of noise (Williamson and Curry, 1984). An additional problem of noise is that it effects the way in which the user speaks (Rollins, 1984; Bond et al, 1986). Under high noise conditions, people tend to increase the volume of their speech. This is known as the Lombard Effect (Lane et al, 1970). By using headphones to feed their speech back to the users, it is possible to reduce this effect.

A second problem for ASR use in the cockpit relates to the rapid acceleration rates of supersonic aircraft. Under very fast acceleration, i.e. greater than +6 Gz, speech articulators in the vocal tract are displaced. This results in marked shifts in the formant frequencies of the speech (Moore and Mc.Kinley, 1986). Finally, pilots of fighter aircraft often experience high levels of vibration at high speed or in battle conditions. Vibration produces a 'shakiness' in the voice, together with an increase in fundamental frequency (Bond et al, 1987).

These effects could be compensated for by having different sets of speech templates recorded under different acceleration conditions. However, although a simple solution, this may not guarantee improved performance; the changes in formant frequencies may simply increase the possibility of confusion. It would be far more useful to incorporate some intelligence in the recognition system to recognise changes in formant and fundamental frequencies, due to acceleration, gravity, or stress, and compensate the recognition algorithm accordingly. There are also certain physiological factors which can affect the pilots voice, e.g. the oxygen mask to which the microphone is attached. Although the effects of the oxygen mask and background noise can be minimised by allowing for them during training, it is difficult to deal with the effects of stress and G force on the voice.

Wilson (1986) makes the point that if some form of filtering was used to compensate for the noise problem, there would be a time delay between spoken utterance and system response. In critical conditions such a delay would be unwelcome. Furthermore, the types of situation which would be critical will result in an increase in psychological stress on the pilot. There has been limited research on the effects of stress on human speech, and this will be reviewed in the chapter fifteen with reference to control room operation.

Limitations of ASR

Promoters of ASR claim that its major advantage over other media, as a means of interacting with computers, is its naturalness. Humans do not need to learn how to speak, it is a skill that they acquire as they grow up. Consequently, it is possible to make absurd claims, such as the following,

> " You will be more inclined under duress to speak (and speak clearly) in your native, natural language than to type, punch buttons, or even speak in an artificial code".
> [Lea,1980]

Currently available ASR devices can only deal with isolated words or short phrases; they require the user to speak in a consistent manner, and to pause between vocabulary units (see chapter thirteen). Furthermore, such systems will use a limited vocabulary. Taken together, these factors question whether ASR systems allow users "to speak... in [their] native language."

Lea (1980) might be suggesting here that, in as much as they are normally occurring words, vocabulary items represent"native, natural language", as opposed to "an artificial code", but the restrictions facing the users of ASR combine to require them to speak in a code of sorts. It is a code in that the vocabulary is necessarily limited and task specific. Using ASR systems requires the user to speak in an 'artificial code' of sorts, in that the user is required to speak slowly and consistently, with pauses using a limited vocabulary. It is important to design a vocabulary which allows users to believe that they are controlling the machine, rather than vice versa. Dialogues can be constructed to elicit one word responses from the user or to call for phrases.

Finally, one could question the underlying assumption of this quotation from Lea (1980): will people be naturally inclined to speak in an emergency? If your car is about to hit an oncoming vehicle will you be more likely to press the brake or say "stop". One major misconception concerning the use of ASR systems, especially in the eyes of manufacturers and advertisers, is the claim that speech is always inherently 'natural'. This is clearly nonsense, as Underwood (1980) points out,

> "...[The] facile argument that, as speech is man's most natural form of communication, spoken communication with computers must be worth doing...is no longer sufficient." [Underwood, 1980]

Speech has evolved to fit the requirements of human communication, and is easily acquired for most people in all cultures. But it does not necessarily follow that speech will remain natural in the context of human computer interaction. Chapter thirteen shows that using ASR is a skill which calls for training and practice in order to develop. People use speech for a number of purposes, and not all of these will be appropriate to speech based interaction with machines (Newell, 1984).

An important distinction needs to be drawn between spontaneous and nonspontaneous speech. In the majority of interhuman communication situations, speech is spontaneous, ie.it is a relatively unconstrained response to variable conversational cues. In HCI speech is nonspontaneous, i.e. it is the constrained response to specific cues. This distinction can extend to cover the use of single words or long phrases. An example of non spontaneous speech would be reading aloud, this can be seen to be analogous in some ways to using an ASR system; the user is required to use only the words which are in the vocabulary.

Chapter four presented a discussion of the limitations of ASR, especially with respect to their lack of linguistic knowledge. ASR can be criticised for being of limited value due to the restrictions placed on the user. However, all computer systems which use some semblance of natural language to operate are run in restricted domains and use a limited subset of 'natural' language (the validity of the term natural language is questioned in chapter ten). This means that any human computer interaction will be limited when compared to human communication. Rather than dismissing ASR as placing undue limitations upon the user, it is necessary to specify what those limitations are and how they can be detrimental to performance.

In terms of language use, ASR can be considered limited on a number of levels. The available range of language is limited by the device's memory space; the way words can be spoken is limited by the recognition process, i.e. isolated versus continuous word recognition; the lack of linguistic and domain knowledge will affect the device's performance; and the rigid structure of the interaction imposed by the syntax will impose limits on the user's style of speaking.

i). Size of Dialogue Unit

The review of commercially available devices in chapter four, showed that ASR could be performed using either isolated or continuous word recognition. In other words, users must either speak word separately with clear spacing between them, or can speak short, connected phrases. Intuitively, one might assume that continuous word recognition (CWR) would always be preferred to isolated word recognition (IWR). However, Lea (1980) has questioned whether CWR would be necessary for most industrial applications. Certainly the applications reviewed above showed that data entry could be performed satisfactorily with IWR. It can be argued that because IWR places constraints upon the users, they will need to concentrate more upon the activity of using ASR and thus, make fewer errors. Also, the use of single words provides clear spacing between utterances and avoids the problem of coarticulation.

Schurick (1986) found that CWR gives faster task completion time than either IWR or manual data entry, on tasks involving the entry of phrases. This finding is supported by Casali et al (1988). Therefore, CWR is superior to IWR for the entry of phrases, such as command strings. When it came to entering words and digits, Schurick (1986) found no difference between IWR, CWR, and manual data entry. So no significant advantage would be gained from using CWR in simple data entry tasks. These findings are further supported by the work of Martin and Welch (1980) reported in chapter seven.

One would expect the size of the dialogue unit (word or phrase) used in the dialogue to depend on whether the device uses IWR or CWR. Cookson (1988), however, reports studies using a CWR device, which allowed users to reply to questions using any number of words. It was found that 91% of all responses used only one word.

The issue of exactly how people speak to machines is crucial to the ideas proposed in this chapter. CWR appears preferable for control room systems, because it allows operators to issue whole command strings. Further, novel strings can be created by combining existing words, and do not require additional enrolment.

ii). Size of Vocabulary

The range of tasks required by ASR in control room systems would necessitate vocabularies to be specifically tailored. For this reason, speaker dependent devices will be preferred. Commercially available speaker dependent devices have a range of available vocabulary of between sixty four and one thousand vocabulary items. The term vocabulary item refers to the unit of speech enrolled as an individual template. It might be a single letter or a number, a word or a short phrase.

Some writers argue that in order for ASR to compete with manual data entry, devices will need several thousand words in their vocabulary (Gould et al. 1983). However, such vocabularies will only be required for the much vaunted speech driven typewriter. The pros and cons of such devices are beyond the scope of this thesis, but it is proposed that they will not offer much benefit to the control room. Furthermore, increasing the size of a devices vocabulary may not make it more 'user friendly'. While the average user of the English language knows around twenty thousand words, the vast majority of the words are not used to any great extent.

Carney (1972) estimates that the fifty commonest words in a persons' vocabulary make up 50% of their verbal output, and that three thousand words make up 92.5%. This means that in everyday uses of language we already limit ourselves to a subset of the language that we know. Different situations require different choices of words, either in the form of jargon or technical forms of expression. Therefore, one needs to ask whether control room systems can be operated with vocabularies as small as between sixty four and a thousand words. This question can be answered by considering the number of words people use to solve problems in very restricted domains, and by considering the number of words used in current applications.

a.) Task Specific Dialogues

Considering that ASR will be used to perform specific tasks in specific domains, the size of the vocabulary required can be gauged by observing how people communicate using speech to solve problems.

Malhotra (1975) used a simulated management information system and recorded subjects queries. Out of 496 recorded queries, she found that subjects used 358 different words. By extrapolating from these figures, he estimated that a vocabulary of between 1000 1500 words would allow people to use the full data base.

Chapanis and his colleagues have been studying the use of language in human computer interaction since the mid 1970's. In one study, Kelly and Chapanis (1977) found that, given, no restrictions on the number of words they could use, subjects employed a task relevant vocabulary of around 300 words. In another study, they found that limiting vocabularies to 300 or 500 words had no effect on problem solving performance compared with unlimited vocabulary. They did find that reported frustration tended to increase as vocabulary size decreased. However, one could argue that frustration was not related to any variation in vocabulary size, but to the subjects knowledge that the experimenter was interfering with their capability to control the system, by imposing certain limits.

Diaper and Shelton (1989) report two studies in which subjects could use natural language to interact with a simulated expert system. In one study, subjects used 468 different words, in the other they used 580. It should be noted that rather than speaking, subjects were typing queries.

Zoltan et al (1982) compared speaking with typing in database enquiry. They found that while speech used 190.7 different words, typing only used 78.8. As well as a difference in the number of words used per condition, they also found a difference in the choice of words per condition. Overall, 840 different words were used by the subjects, 642 in the speech condition and 373 in the typing condition. This suggests that speech is more verbose than typing as a means of database enquiry. The verbosity of speech over typing is further shown by Ford et al (1980).

Hauptmann and Rudnicky (1988) compared subjects typing or speaking to people or speaking to computers on an electronic mail system. They found that while both groups using speech employed a similar number of different words per session (36.65 between people, and 32.7 between people and computer), the typing group used 23.75. The difference was significant ($p<0.01$), suggesting that subjects try to be more concise when typing, possibly in an effort to save their workload.

Carbonell and Hayes (1983) distinguish several types of 'sentential extragrammaticality' which can occur in human computer dialogues. They found that users often employed a form of 'computerese', in which function words, articles etc were omitted from the input. This was suggested to be a way of making the input easier for the computer to understand.

The results of all these studies combine to suggest that whenever subjects are required to use natural language to perform computer based tasks, they tend to employ a very limited subset of language. The size of the vocabulary used is well within the range for commercially available ASR devices.

b.) Current Applications

Current applications reported in this chapter use, on average, 40 separate words. Nye (1980) estimated that 80% of all IWR applications of ASR use 75 words, and 66% of CWR applications use 200 words. The demonstration system reported in chapter nine required a vocabulary of 37 words, but it was clear that this figure could be extended to include more place names.

The studies and applications reported suggest that vocabulary size need not limit the industrial use of ASR. Most applications in control room systems would require a small vocabulary of command words and additional place and object identifiers. Such a vocabulary could be implemented on most commercially available devices.

iii). Syntax Limitations

With limited vocabularies comes a limitation on how units of the vocabulary can be combined into meaningful phrases. Some ASR devices employ a type of syntax which specifies which units are permitted to follow each other. Therefore, once a word has been recognised the device will only attempt to match the next word from a limited subset. This reduces the search space and ought to reduce confusion errors. But problems can occur, most catastrophically when the first word is incorrect and the syntax could take the user to completely incorrect points. The use of syntax also puts further work on the user, who must know which will be an appropriate word to speak next.

Thompson (1980) complied a corpus of 1615 queries made by users to a database . Of these, 446 were erroneous. Even if the 161 vocabulary errors are dismissed as the result of typographical mistakes, this still leaves 285 errors. These are due to either incomplete input strings or syntax errors. Reisner (1977) proposed that syntax errors in database query languages often occurred when users try to couch their requests in syntax related to English rather than the query language. This suggests that user formulate commands in English (or their native language) and then translate them into the query language. Syntax errors could be reduced by employing a syntax which is similar to that used in English.

Diaper and Shelton (1989) classified 90% of the 675 queries to their 'expert system' into five phrase types. The most common type used a noun phrase followed by a verb phrase. The other four were noun phrase; verb phrase; auxiliary verb phrase/ noun phrase/ verb phrase; adverbial phrase/ noun phrase/ verb phrase. These results complement the findings from the vocabulary size section, and suggest that subjects tend to use simple task specific vocabularies in short, simple constructions.

iv). Knowledge Limitations

It has been argued that the apparent limitations of vocabulary size and syntax need not restrict the use of ASR. In order to use language efficiently, the user must have some knowledge of that language. Chapter four showed that the use of knowledge components in ASR was very limited.

People use different aspects knowledge for specific tasks. Dumais and Landauer (1982) asked subjects to define words. They found that subjects rarely used negatives or opposites, but began their definition with a superordinate followed by either an example of the word or listing some of its attributes. This shows that one cannot simply speak of using knowledge of language per se, but must consider what function the knowledge will perform.

In any application of ASR, words will refer to actions and objects. At one level of analysis, if an ASR device recognises a word and sends a command to the host computer to perform an action, it has "understood" that word. The knowledge is contained in a rule <if X then Y>.

> "...[W]hen told that someone or something's understanding is limited, people, on the basis of their experience with language, assume that the shortcomings are in vocabulary or in understanding concepts."
> [Leiser, 1989]

Background noise will interfere with ASR performance. There have been several suggestions as to how to reduce the effects of this problem, but it needs careful consideration in any application design project.

Conclusions and Guidelines for Application Selection

From the discussion so far it can be seen that ASR is potentially useful in some situations. However there is still a reluctance on the part of industry to commit itself to installing ASR equipment. Figures from market analysts in the US reflect this, along with the optimistic predictions of expanding markets. This is illustrated by figure 4.2.

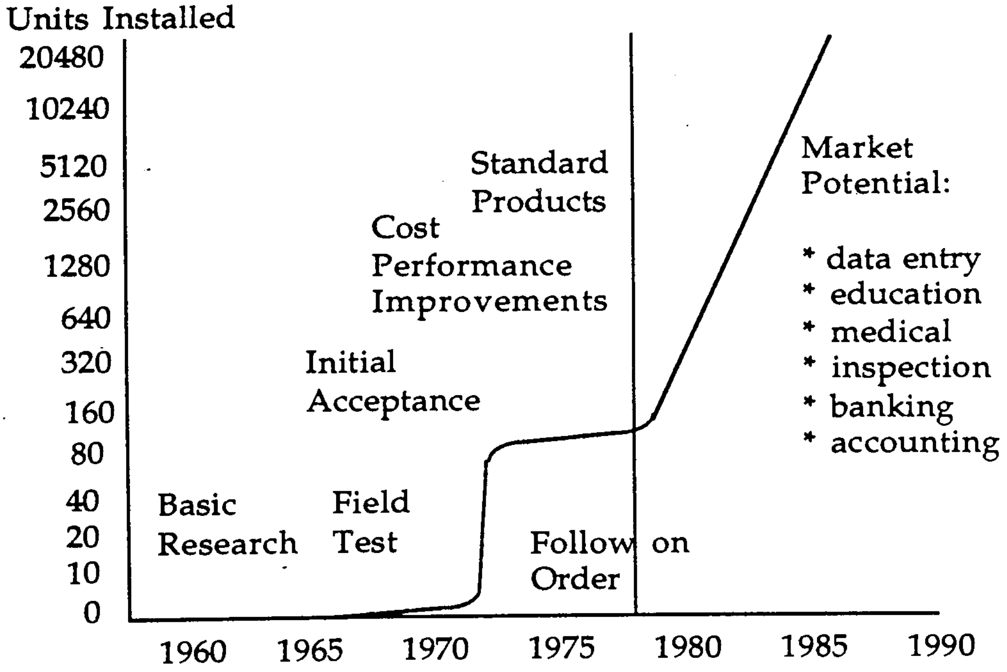


Figure 6.2: Market Forecast of ASR Potential [From Nye, 1980]

International Resource Development suggested that ASR would reach a market worth $4 billion by 1982. Their own figures for 1984 showed that the US market alone had only reached $115 million, and the bulk of sales were for speech synthesis and speaker identification devices (speaker identification uses templates of a persons voice to identify them, it is used for security purposes).

In the U.K., automatic speech recognition is taking longer to achieve any sizable market. Noyes and Frankish (1986) found two applications in the U.K., at the hydrography department (see Starr et al. 1988) in the Ministry of Defence, Taunton, and by Clyde Surveys Ltd., Maidenhead for map making. In 1987 Austin Rover Group introduced ASR for online vehicle inspection (Anderson and Gill, 1987), and in 1988 ASR was introduced for vehicle inspection at Jaguar cars (Cassford, 1988) and Caterpiller Trucks (Peckham, 1989). There has been much work carried out by the Royal Signals Research Establishment at Malvern, and by the Royal Aircraft Establishment at Farnborough into the use of ASR in aviation. It is not known if any applications exist outside of the laboratory. This review, although possibly not extensive, shows that ASR is gradually finding a niche in British industry, but progress is painfully slow.

One major obstacle that ASR meets time and again, is the comparison with human speech processing. This is problematic a two counts: it leads to excessive, often unwarranted, optimism concerning what ASR can do and where it can be used, and this in turn leads to marked disillusionment when first meeting ASR (Doddington, 1980). This pattern is predicted in the Introduction, with peoples' experience of speech technology mainly derived from science fiction. However, the problem is further confounded by the fact that ASR relies on a skill we take for granted. When one models human skills, people expect the performance to be of a similar standard to humans. In this respect ASR technology is still very primitive. Manufacturers do not help this problem by quoting performance scores derived from the pristine conditions of the development laboratory (see chapter five).

Given that ASR is a poor second best to human speech recognition ability, one might ask what use installing ASR equipment will have, when a human operator could do the process much more easily? This question raises three points:

i.) if one refers back to the example of the baggage sorting system (Backer, 1984), one realises that ASR can be used to improve the performance of operators on some tasks, by reducing their cognitive workload;

ii.) from the consideration of ASR to any application mentioned above, it is clear that ASR is not a panacea to all applications problems, and consequently should not be over rated;

iii.) the speech handling capacity of all commercial and research ASR systems is much lower than that of humans, so one must pay particular attention to the human factors of the interaction to alleviate mismatch.

This chapter has shown that, despite the apparent limitations of the existing technology, ASR can be successfully applied to a number of situations. One thing the applications have in common is that they require simple input messages. Furthermore, all the applications considered here can be considered as finite, linear tasks; by this I mean that the tasks in which ASR are best used, at present, are ones which do not allow the user to deviate from an already established routine, e.g. in a procedure for fault finding. Finally, the applications considered here share a requirement for vocabularies of the 'either/or' format, i.e. they do not use command sequences which can operate in serial tasks, such as controlling the flow of liquid, but only in tasks which require either one discrete command or another.

In the applications where ASR has been successfully used, benefits have resulted from one or more of the items listed below. In addition to these, specific applications may yield specific benefits, such as in the example of baggage handling (Backer, 1984). The introduction of ASR increased throughput, and as a result eliminated the time consuming problem of baggage 'bunching' on the conveyor belts.

* Hands/eyes free
* Elimination of hand recording of data, and associated transcription errors
* Ease of use
* Reduction of eye strain in long term data entry
* Limited requirements for user training
* Mobility of user

* Naturalness of command language

* Source data capture improved

* Simple hardware interfaces

* Rapid data entry

* Flexible vocabulary, rather than being 'hard wired' like a keyboard, it is possible to substitute any word for the desired command. This leads to,

* Possibilities to tailor command vocabulary to specific situations

* 'Infinite' soft key facility, i.e. one spoken command substitutes for several key punches

* Reduced desk space requirements

There are a number of possible applications, suggested by Kelway (1988), which have yet to be explored. These include the use of ASR in wet, damp, or windy conditions; unstable locations, such as ladders or platforms; in confined areas, dusty, dirty , and greasy locations. This suggests that ASR could be viable in the inspection of transmission lines, or power station plant. Such task domains lie outside the scope of this thesis, but deserve further investigation.

The limitations of ASR do not seriously diminish its potential in the control room, and are outweighed by the advantages derived from the review of applications in this chapter. ASR presents a viable means of interacting with machines and computers in a range of contexts. Chapters eight and nine present the results of a study into the potential application of ASR in a control room context, specifically in a grid control room.

CHAPTER SEVEN

ASR versus MANUAL INPUT DEVICES

Applications of ASR often show improvements in task performance, over previous methods. It is difficult to propose whether such improvement is the result of some inherent advantage of ASR over manual control or whether the introduction of ASR requires a redesign of the basic task which, of itself, aids performance.
This chapter discusses the issues involved in comparing ASR with manual input devices. The advantages of one medium over the other are shown to be largely task specific. Consequently, it is argued that careful consideration needs to be given to the type of task which ASR is used for. From the discussion, and from Study 1, it is possible to draw a set of principles to inform application selection and design decisions.

Introduction

The discussion of existing applications for ASR, in chapter six, illustrates the wide scope for potential use. The most common application is as a means of data entry, especially in situations where users have their hands occupied. Further, ASR can offer an additional channel of communication in complex domains, such as the fighter aircraft cockpit. Both of these points offer the potential to reduce problems in human computer interaction in control room systems.

When manual data entry is used it often interferes with operator efficiency. When the operator has very limited keyboard skills, the task of manual data entry can be time consuming (taking up almost 1/3 of total task time n terms of searching for, and striking, keys) and prone to error. Also, if the manually collected data has to be analysed and then acted upon, errors may not be detected until the process is complete. Voice data entry can help to reduce transcription errors and eliminate paper data.

However, ASR should not be regarded as a panacea for all HCI problems. Tasks should be selected which will benefit form the use of ASR, rather than using it as a mere add on. The issues involved with the selection of appropriate tasks are discussed in detail in chapter six. ASR can have advantages over manual control in several situations. But direct comparison between ASR and manual control is often problematic.

ASR as an input medium for Process Control Operation

Carey (1985) proposed that input devices for human computer interaction could be investigated with respect to their performance on five generic input operations, viz.

i.) Specify object.
ii.) Specify location.
iii.) Enter numerical value.
iv.) Specify required action.
v.) Enter text.

He suggests that process control will be primarily concerned with the first three operations. With the exception of text entry, virtually all manual input devices could satisfy this range of input goals.

The comparison of input devices is made difficult by their different characteristics, and demands upon the operator. However, two measures which are commonly used, relate to speed and accuracy of use. It is these measures which will be used in the following discussion. While Carey (1985) proposes that process control requires input devices which allow operators to specify object and actions, and to enter numerical data, many control operations also call for text entry, for example, in the form of operation logging. For this reason, the following discussion conflates textual and numerical data entry under one heading.

The most common medium for inputting alphanumeric data is the keyboard. This will be compared with ASR, in terms of a number of studies reviewed. Modern VDU based control room displays, using high resolution graphics, allow plant objects to displayed in terms of their relationship to each other, often in the form of plant diagrams. Specifying the location on such diagrams, therefore, simply requires the operator to point to the desired object. In his review of different input devices, Carey (1985) found that the best media for this task were touch screens and function keyboards. Study One reports the comparison of ASR with a function keyboard on a process control task. The operator was required to specify a particular object in order to receive information concerning its performance level.

It could be argued that specifying an action results naturally from specifying the object, as in the simulation used in Study One. However, it may be necessary to specify different actions for several objects. One could imagine the use of a pull down, or pop up menu display, or a set of function keys, or a spoken or typed command string to perform this operation. Research has yet to be completed comparing ASR with these other forms of action specification, although the conclusions from the discussion of alphanumeric data entry has a strong bearing on this point.

Keyboards vs. ASR for Alphanumeric Data Entry

Everyday experiences of speaking and typing would lead us to conclude that it is easier, and faster, to speak a sentence than to write. Such advantages have formed the basis of claims for using ASR, but it is important to ask whether speech is always faster than typing. In an oft cited study, Welch (1977) compared an isolated word recogniser, a conventional keyboard, and a menu data entry system on two types of data entry tasks. Simple data entry consisted of inputting strings of three to ten characters. Performance using the keyboard was fastest and most accurate in this condition.

Complex data entry consisted of composing flight data control messages. Spoken data entry was shown to be faster than the other media. It was also subject to fewest user errors in the 'simple' condition. This was due to the fact that ASR allowed users to keep their eyes on the list of character strings to be read in, and so produce fewer reading errors. However, the limited accuracy of the ASR device led to device errors which required correction.

These results support two conclusions. A keyboard will be faster and more accurate than ASR for simple data entry tasks. Spoken data entry is more suitable for complex data entry and command tasks. But these conclusions need to be treated with some caution. These conclusions are further supported by Connolly (1979). He found that entry of Air Traffic Control messages was faster using speech than a keyboard. However, if the message consisted solely of digits, then the keyboard produced faster performance. The results from these studies need to be considered in some detail, in order to draw out design principles, and can be divided into speed and accuracy measures.

Speeds of Data Entry for ASR and Keyboard

Trained typists can enter between 1.6 - 2.5 words per second, while untrained typists manage 0.2 - 0.4 words per second (Lea, 1980). On a small vocabulary application of isolated word ASR, Martin (1977) suggests words are entered at a rate of 0.5 - 2 per second. It is faster to speak a word than to type it it. Poock (1982) has extended this common sense argument to cover commands which use several words. He was able to show a significant superiority of spoken, to keyboard, data entry. However, in his study, multiword commands were enrolled as single templates. This meant that each command could be spoken as a single item, e.g. "forward message", but had to be typed as a string of letters.

In a study comparing keyboard use with several possible types of speech data entry, Schurick (1986) found several interesting results. She divided speech into entry using phrases, individual words, and spelled words. The latter group required subjects to spell each word, letter by letter, either using the conventional alphabet (orthographic), or the ICAO alphabet (phonetic). Schurick (1986) found that there was no significant difference between the accuracy of the keyboard and phrase level speech (and both were more accurate than the other groups), data entry using phrase level speech was significantly faster than the other groups. However, comparisons of spoken and manual data entry should be regarded warily. While they can be used to perform the same tasks, with the same vocabulary, the units of comparison between ASR and manual data entry are different. For keyboards, it is the single keystroke; for ASR it is a word or a whole phrase (McCauley, 1984).

The majority of keyboard users in control room systems have very low levels of typing skills; typing is often performed using a single finger. Although trained typists can easily navigate around a querty keyboard, untrained users follow a "search and strike" approach. For this reason, it is customary to design control room systems which are run using minimal key strokes, or dedicated function keyboards.

Damper (1988) argues that the point at issue in comparing ASR with manual devices, is whether it is faster to type or to speak a single unit.
(Mullins, 1988) also recognises the importance of this issue, defining the way in which users will organise and control rapid sequences of movements, either in typing or speaking, as "A fundamental concern of input technology".

Practised typists take between 100-200ms to locate and depress a key (Hershman and Hillix, 1965). Untrained typists perform the same operation in 1000ms (Devoe, 1967). Skilled users of ASR took between 400-800ms to enter words of the ICAO alphabet into an isolated word recogniser, with an additional processing of 200ms per word and a 200ms pause between words. Thus, a single word would take up to 1200ms to enter. This would give a rate of 50 words per minute (see Martin, 1977 above), which compares favourably with typing words letter by letter.

It is important to note that Damper's (1988) figures are based upon comparison of isolated word ASR and keyboards. Many devices on the market now support connected word ASR. This means that users do not have to pause between words, or rather that the required pause is negligible. Further, processing times are increasing each year and so it would be erroneous to dismiss a technology on the basis of out dated examples. Taking these points into consideration, Damper's original 1200ms can be reduced to around 800ms. This is still slower than the speed with which a trained typist can type individual letters, but offers a potentially faster medium of phrase level entry.

The comparison of entry speed can be further investigated from the theoretical perspective proposed by Sternberg et al (1978, 1980). Their motor programming hypothesis suggests that programs of physical movements are represented at a cognitive, as well as a physical, level.

Mullins (1988) proposes that intention is an essential component of movement, and takes the form of an action plan. This proposal can be tested very simply. If the meaningfulness of stimuli will effect processing times, then some form of intervening cognitive processing can be assumed to be taking place.

Mullins (1988) presented subjects with a list of one to five stimulus items, "Digits, syllables, or words in the speech conditions; one finger or alternating fingers in the keyboard entry conditions." After an interval of 2.5s, subjects were required to repeat the stimulus as quickly and fluently as possible.

When digits and syllables were used in the speech condition, and when one finger tapping was used in the keyboard condition, the motor program hypothesis was supported, ie. the reaction time increased with the length of the response string. However, when words were combined into meaningful phrases, reaction time functions were flat. Even though the response output consisted of single words, the result of imbuing the combination with meaning led to faster reaction times.

This suggests that the meaningfulness of a response will effect the organisation of movement, and decrease reaction time and response variability. Thus, speed of data entry is increased by organising it into meaningful contexts. This notion can be used to reconsider the findings of Welch (1977): that simple data entry should be carried out using a keyboard, and complex data entry using speech.

For 'simple' data entry tasks, if we assume that the task of entering data on a numeric keypad will be carried out by most subjects using a single finger, Mullins (1988) proposed a flat reaction time function, owing to the possibility of formulating an action plan for the sequence. The space covered by a ten figure numeric keypad is quite small, and so key presses will require similar finger movements. The difference will lie in finding the appropriate key for each number. On the other hand, entering digits by speech will result in a reaction time proportional to the length of the string (in agreement with the motor command hypothesis).

In the 'complex' data entry condition, typing a series of words will require alternate fingers to be used. Mullins (1988) showed that two finger typing produced an increasing reaction time proportional to the length of the string to be input. Speaking a meaningful sequence of words, however, yields a flat reaction time function. Therefore, Welch's (1977) results can be seen to conform to Mullin's (1988) action plan hypothesis, and can be used to generate the following proposals:

i.) Data entry requiring single key presses, e.g. digits or function keyboard use, will allow users to formulate an action plan for their responses. This will result in fast, accurate data entry.

ii.) Data entry requiring meaningful commands will be best performed using speech. Users will be able to formulate an action plan and perform fast, accurate data entry.

Thus, careful consideration needs to be given to the selection of appropriate applications for ASR, in terms of task requirements. For instance, the task of command entry, which could be proposed as a common requirement in control rooms, is well suited to the use of ASR

Accuracy of ASR and Keyboard

Speed of entry is not the only criterion for judging input media. Accuracy is often a more critical variable than speed, especially in control room systems. In many current applications of ASR it is not important that commands are recognised as quickly as possible, or even that all word are recognised correctly. Chapter twelve discusses the issues surrounding error correction. It is necessary to decide whether the time the user must spend in correcting errors can be reduced by using a more accurate device, which would be more expensive, both economically and computationally, or whether error correction can be incorporated in the main task of using ASR without too much difficulty.

By using some form of error correction, Welch (1977) found speech data entry to be as accurate as a keyboard. However, this conclusion overlooks a very important distinction in the occurrence of errors in ASR and keyboard use. The keyboard as a 'standardising' effect on input (Knight and Peckham, 1984). This means that providing the desired keys are depressed to the correct extent, data entry can be carried out in a number of ways, eg. using a single finger or a pen. However, speech is inherently variable (see chapter three),which means that data entry must be carried out consistently, to achieve adequate matching to the stored representations. This is reflected in the errors that are produced, and suggests that ASR will inevitably be less accurate than typing.

Current ASR recognition accuracy is in the region of 97% (although, as we discussed in chapter five, this is not necessarily a clear indication of performance). The study by Schurick (1986), reported above, found no significant difference between the accuracy of keyboard use (99.17%) and phrase level speech data entry (95.35%). The following studies report further comparisons of ASR and manual data entry. If ASR can be shown to have an acceptable level of accuracy, when compared with a keyboard, then it deserves serious consideration as a computer input technology.

Studies Comparing ASR with Keyboards

Even the most accurate ASR devices will have some possiblity of errors occurring. A study by Casali et al. (1988) shows that the accuracy of an ASR device has a significant effect on task completion time and user's acceptability ratings. However, their measure of task completion time is somewhat dubious, in that it is intrinsically linked to device accuracy; the less accurate device required users to spell out misrecognised words, letter by letter. Thus, the time measured could have been an artifact of the experiment, rather than an indication of the effect of accuracy on performance.

Voice input was compared with function keyboard and qwerty keyboard in a study by Mutschler (1982). He found marked differences in the accuracy with which the devices were used to perform the task of inputting a set of words. However, the vocabulary for the study consisted of German Male Christian Names, combined into lists of ten, thirty, or ninety words. It could be argued that such a vocabulary does not provide a suitable basis for conclusions of ASR use, as the words are not used in any meaningful context but brought together in an arbitrary list. A task specific vocabulary would give the words not only individual meanings, but a semantic and functional context (see chapter ten).

We can extrapolate form Mutschler's (1982) results to discuss why a difference in accuracy exists between the three media studied in terms of different vocabulary sizes.

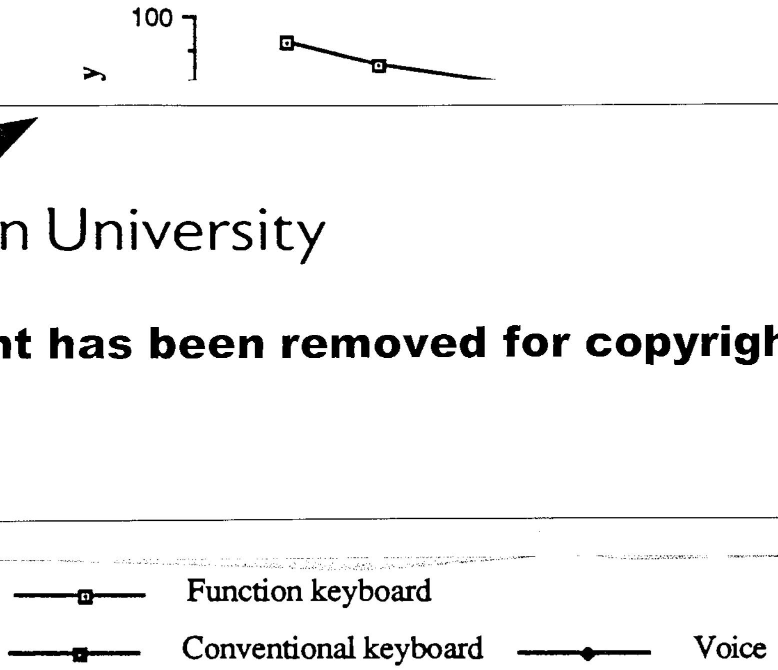


Figure 7.1: Graph of Mutschler's (1982) Results

i) Large Vocabularies

With large vocabularies (90 words and over), a function keyboard imposes a search task upon the operator. The scanning demands required to find the right key will be sufficient to slow down the interaction, and also, possibly, to lead to some errors. Further, a function keyboard for a ninety word vocabulary will take up a great deal of space on the control room desk. ASR could be considered analogous to a function keyboard, in that a single command could be substituted for several keypresses on a conventional keyboard.

Poock (1980) points out that the labels on function keys can serve as useful memory aids to the operator, to help facilitate recognition of appropriate commands words. Such an aid is not available to users of ASR, which will inevitably have fewer cues to the range of available inputs. Obviously one could display all possible words on the screen, but for large vocabularies this will take up too much space, and will introduce the problems of scanning found in function keyboards.

It is common for psychologists to distinguish between recognition and recall memory. Recognition occurs when a stimulus, e.g. a face, matches a previously learnt item. Recall, by contrast, involves the retrieval of the previously learnt item from a memory store. In the case of ASR and function keyboards, the items to be remembered are the words in the vocabulary. A user can recognise words written on the function keyboard's keys, but has no opportunity to do this with ASR. But even if the user could recognise the words, he will still have to recall their meaning, e.g. he might be able to search for, find and recognise the word "import" on a function keyboard and still be unable to recall what the command actually does.

One would imagine that operators construct their commands by first deciding what needs to be done and then deciding on the appropriate command. In the case of using a function keyboard this might be the pressing of a sequence of keys, in using ASR it might be the speaking of a particular command. ASR will normally be used to enter data or issue commands in control room systems. Providing the commands are issued in an easily construed task context, using a 'habitable' vocabulary, recall should not be a major problem (see chapter ten).

ii) Small to Medium Vocabularies

With medium and small vocabularies, the function keyboard will be easier to use than the qwerty keyboard. It will require fewer keystrokes and produce less possibility of errors. ASR is analogous to a function keyboard, but has the advantage of being more versatile. While a function keyboard is built for a specific set of words, the ASR device will pair any word with a particular ascii code. The device can therefore be tailored to suit the user simply by retraining the vocabulary.

The qwerty keyboard was assumed to have a constant level of accuracy, proportional to the users' typing skills. Simple alphanumeric strings result in the qwerty keyboard being used like a function keyboard. If command words were to be used, they would have to be typed letter by letter. This is both time consuming and subject to errors for untrained or occasional users.

It has been common practice to abbreviate command words to make typing less arduous. But this results in problems of remembering the artificial codes used (McMillan and Moran, 1985). ASR uses normal words, and will not be subject to such problems.

Conclusions

Therefore, ASR ought to be easier for the operator to use in a number of circumstances. Mutschler's (1982) results do not measure ease of use, but rather accuracy of the device. It is clear that ASR's performance decreases dramatically with an increase in vocabulary size (see chapter five). Although there is much ongoing research to develop large vocabulary ASR devices, the technology is still best able to support small to medium size vocabularies. As the discussion on industrial applications in chapter six suggested, ASR can function very well on a limited vocabulary.

One of the advantages cited for ASR, in chapter six, was that it could be used as an infinitely extensible function keyboard emulator. In the Mutschler (1982) study above, ASR compares favourably to the function keyboard on small to medium size vocabularies, which is the probable vocabulary size for industrial applications. Whilst one could argue that ASR is preferable to a function keyboard, in terms of desk space required, further research is necessary to compare these media on real tasks (as opposed to the entry of German names used by Mutschler, 1982).

The findings of Welch (1977) suggest that ASR will be suited to 'complex' tasks. Humans are able to convey complicated information and solve problems more easily using speech than other communication media (Chapanis, 1977), and this argument has been used extensively by promoters of ASR. Yet there have been few studies reported which investigate the role of ASR in 'complex' tasks. No doubt this is due to the problems of defining acceptable tasks which allow some degree of complexity for the subject and can produce measurable results. All too often task complexity leads to overwhelmingly complex data, which defeat attempts at interpretation. However, two studies reported below investigate the use of ASR and function keyboards on tasks which can be reasonably be defined as 'complex'.

ASR vs. Function Keyboards for Object Specification

Brandau (1982) compared the use of an Heuristics 7000 isolated word ASR device with a dedicated function keyboard on a target acquisition and identification task. A 'target' appeared somewhere on a VDU. Subjects had to issue the command

"Acquire", followed by the number of the quadrant in which the 'target' appeared. If these inputs were accepted, then a cursor appeared which subjects positioned using a joystick, and issued the command "Enter" to fix on the 'target'. Finally, they had to identify the 'target' as being "Friend. Foe, Neutral, or Unknown".

Brandau (1982) found that ASR produced significantly slower and less accurate responses than function keyboard use. He noted that the differences were partly attributable to the time delay in the ASR device. Using isolated speech meant that, although the study compared the media in terms of roughly equal length input units, ASR will naturally be slower than the keyboard. Rather than being able to speak whole phrases, such as strings of digits, the user will have to pause between each unit. A second explanation of the results related to the inconsistency in the feedback provided to the user. Some commands did not receive any obvious feedback. In these situations, the subjects in the ASR group were uncertain as to the accuracy of the ASR device's response, and paused to check the display.

Consequently, a main conclusion from this study is that when recognition accuracy is uncertain, and when the cost of misrecognitions is high, users' speed on tasks using ASR will be affected. These points can be dealt with by providing the user with adequate feedback (see chapter eleven), and a means of correcting device errors (see chapter twelve). It should also be noted that Brandau's (1982) study was based on research into avionics, and may not transfer to process control tasks. The following study compares the use of ASR and a function keyboard on a simple process control task.

Study One

This study was carried out to compare the use of either a function keyboard, or isolated word ASR to run a process control simulation.
Subjects were required to call up displays of the outputs from different plant units, in order to monitor plant performance. Deviation from the base level was corrected by resetting the output from a particular plant unit (this task is explained in more detail below).

Ten Undergraduate students from Aston University served as subjects for this study (six male and four female). They were assigned to one of two groups, relating to mode of input used: i. ASR; ii. Function keyboard.

The simulation was run using a BBC Model B microcomputer with a 6502 second processor, to drive a low resolution RGB colour display. The software had been written for previous research in the Department, to use a custom built dedicated function keyboard (see figure 7.2). The software was modified to incorporate the use of speech. ASR was performed using a Votan VPC2000, voice processing card in an Opus PC II, interfaced using the voice programming language provided with the Votan Inc. A separate piece of software contained a script for the occurrence of plant faults, which occurred every thirty seconds or so, and ran for twenty minutes.

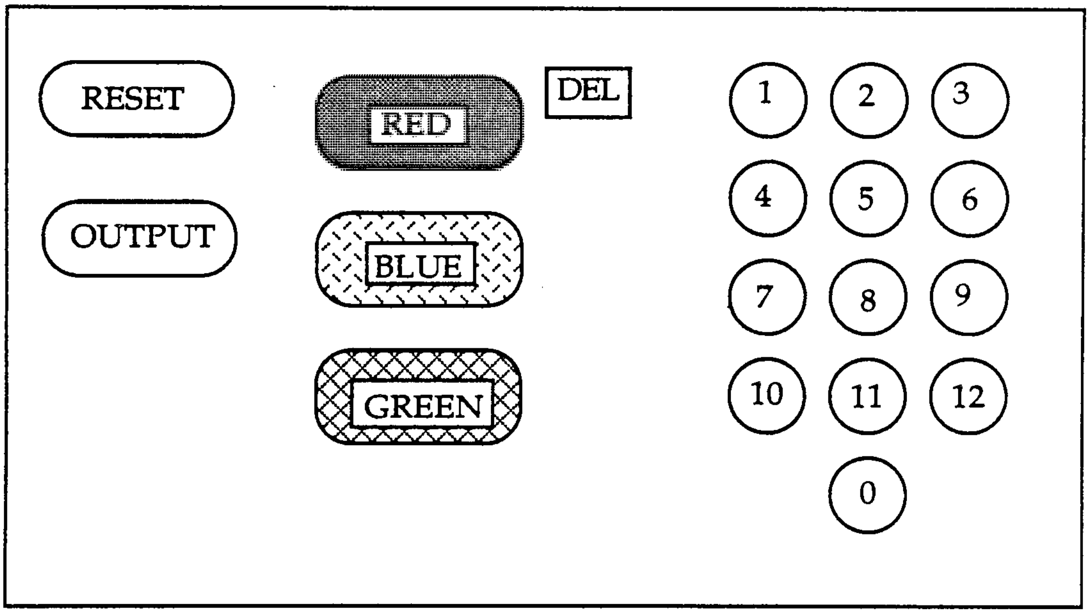


Figure 7.2: Function Keyboard used in Study One

[Note, the 'colour keys' on the keyboard used were coloured appropriately, rather than labelled. The diagram illustrates their spatial position and notes their colour]

Figure 7.3 illustrates the plant diagram, which appeared on the screen. The display is a schematic diagram of a process control plant. The lowest level units are numbered 0-8. These represent 'plant units'. In other words, areas in which the process occurs.

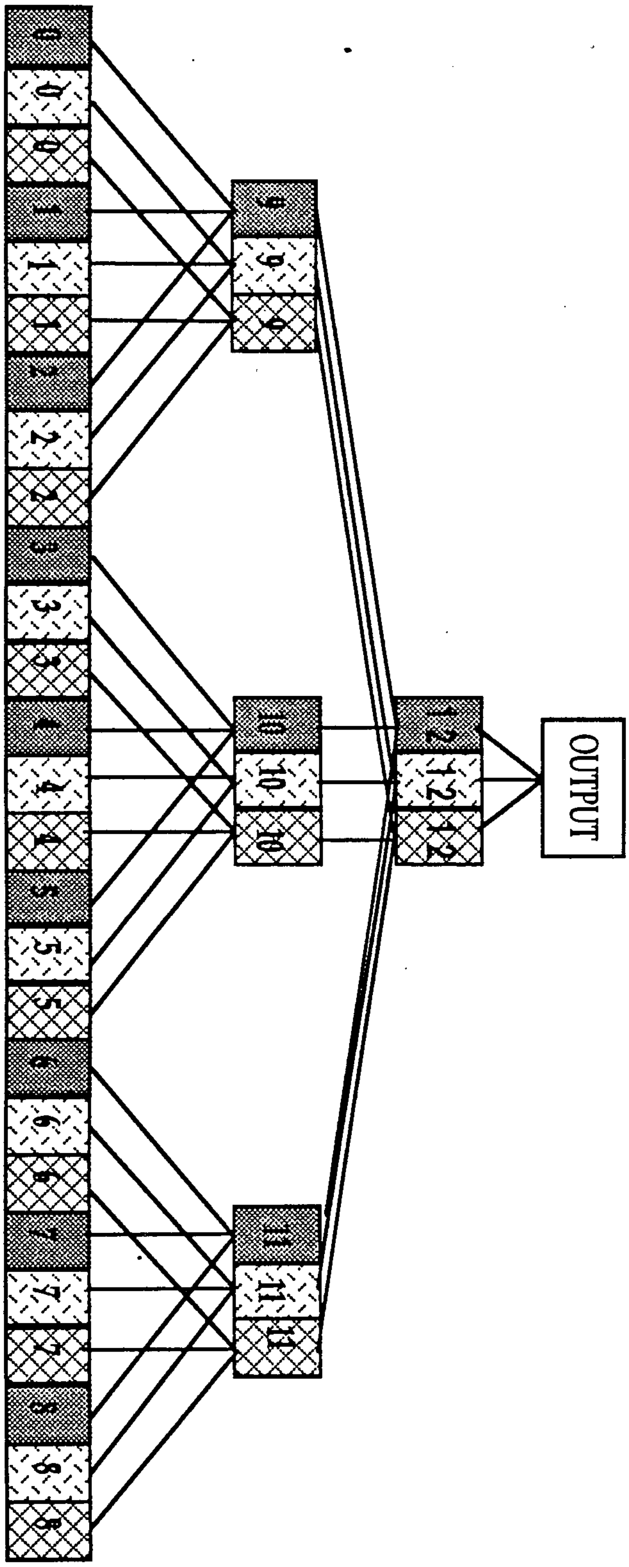
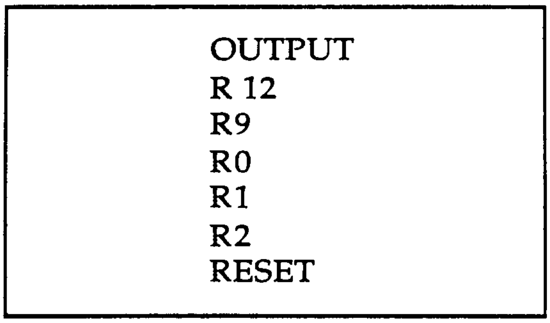


Figure 7.3: Schematic Diagram of Process Plant

The next level, 9-11, represent 'plant monitors'. Groups of plant three units are monitored by one 'monitor unit'. The grouping is governed by the colour coding and sequential order of unit numbers used in the diagram, such that 'plant units' Red 0, Red 1 and Red 2 are monitored by 'plant monitor' Red 9. The next level, 12, represents 'branch monitors'. Each of the three colour coded branches in the plant diagram in monitored by one 'branch monitor'. Each unit and monitor is, therefore, identified by its colour and number. At the top of the diagram is an 'output monitor', showing overall changes in plant performance. The subjects task was to keep the plant output as low as possible. This was achieved by monitoring plant activity, at the output level, until a deviation from the base value was noticed. The subject then traced the deviation through the monitors to the 'plant unit'. Deviations could only be corrected at the 'plant unit' level. Correction took the form of issuing a command to Reset the unit.

Above the plant diagram (not included in figure 7.3), is an 'alarm panel'. Deviations of more than two hundred points triggered an alarm. Below the 'plant diagram' were the 'Operation Log' (see figure 7.4) and the 'unit/monitor display' (see figure 7.5).



```
OUTPUT
R 12
R9
R0
R1
R2
RESET
```

Figure 7.4: Operation Log

The 'Operation Log' echoes the subjects commands to provide user feedback. In this example, the user has traced a deviation to a 'plant unit' and reset it.

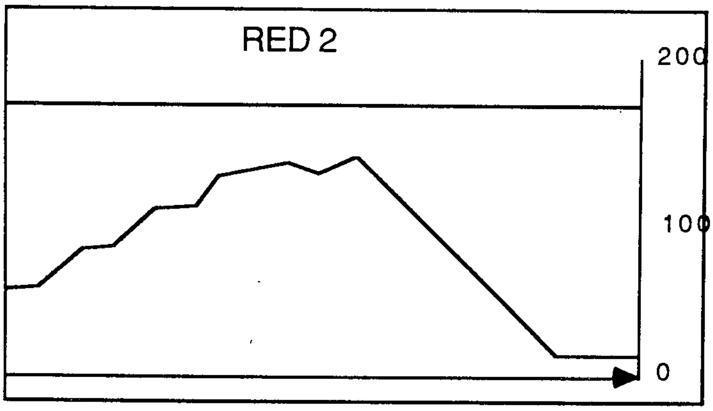


Figure 7.5: Unit / Monitor Display

The 'unit / monitor display' shows the output of the selected unit against time. In figure 7.5, the unit has recently been reset back to the base level.

Subjects on both groups received an explanation of the process, followed by a brief demonstration by the experimenter. The demonstration used one or other of the input media, depending on the subjects group. Following this, subjects in the ASR group enrolled the device using three passes per vocabulary item. The vocabulary was simply a verbalisation of the keyboard panel. After a five minute practice session, during which time performance accuracy measures were recorded, subjects spent twenty minutes controlling the process using the function keyboard or ASR (It was not possible, at the time the experiment was conducted, to record recognition accuracy scores during ASR performance of the experimental task). The output of the plant was recorded by the computer at set times. These output measures were averaged every minute, to give a set of twenty plant output scores. These are the scores used in the comparison of the two input media. It was felt that one needed a measure which would reflect system performance rather than device use, in order to produce a meaninigful comparison. After finishing the task, subjects were asked to complete a short self report questionnaire. Answers took the form of responses on a rating scale.

Results

An Analysis of Variance was calculated to compare the performance of the two input media. It had also been noted that the male subjects were achieving slightly higher recognition scores than the female subjects, during the practice session. It has been noted that ASR seems to be designed to hamper the performance of women users, ansd so the ANOVA was also used to investigate sex differences. Therefore, the variables included in the ANOVA were input device and sex, together with the plant output measure data.

Source of Variation	d.f.	Sum of Squares	F	p
Input device	1	22540.01	47.769	0.00001
Sex	1	2171.704	4.602	0.0476
Input device / Sex	1	1471.1	3.118	0.0965
Error	16	471.856		
Time	19	619.463	6.839	0.00001
Input device / Time	19	375.362	4.144	0.00001
Sex / Time	19	86.366	0.954	0.5168
Input device / Sex / Time	19	150.431	1.661	0.0419
Error	304	90.574		

Figure 7.6: Table for ANOVA of results from Study One.

The analysis yields a number of interesting results. The first set of results concerns the difference between male and female subjects. The was a significant difference between men and women overall [F(1,16) = 4.602, p<0.05], which was borne out by the significant difference in performance using the two input media [F (1,16) = 3.118, p<0.01]. Further analysis, using a Tukey test revealed that while there was no significant difference between sex and function keybaord (p< 0.792), there was a significant difference between sex and speech input (p< 0.1), with male subjects achieving higher performance scores than female subjects.

The second result concerned the overall performance using the two input media. This result was found to be highly significant [F(1,16)= 47.769, p<0.00001].

The third result concerns the amount of performance variation with time. Performance scores for all subjects varied significantly over time [F (1,304) = 6.839, p< 0.00001]. Again, differences were found between male and female subjects [F (1,304) = 86.366, p < 0.5], although the level of significance is quite low. It was found that performance scores varied with the type of input device used [F (1, 304) = 4.144, p < 0.00001].

Figure 7.7 shows a graph of the variations in output level over the experimental period for the two groups. It is clear that ASR produced not only more variation in control, but also greater average deviation from the base level than the function keyboard group.

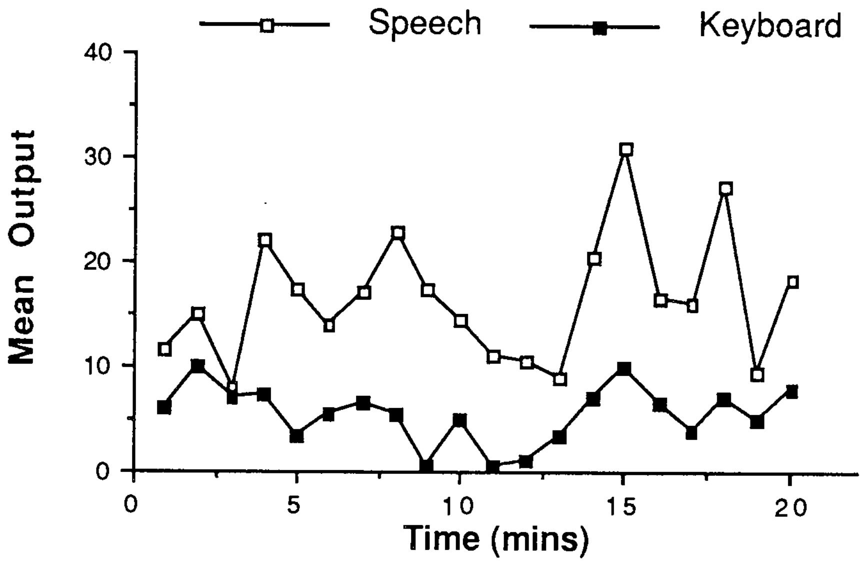


Figure 7.7: Graph of Mean Output Variation for Speech and Function Keyboard Control

This difference in performance between the two groups can be broken down into separate categories, as shown in figure 7.8.

measure	function keyboard	speech input	
No. commands per minute	20	12	
Mean time to rectify faults, per five minutes	39s	72.25s	
No. faults missed	10	35	
No. Alarms	0	1	
Accuracy	99%	96%	(pretest)

Figure7.8: Table of Performance Measures

Subjects using ASR to control the process took longer to rectify faults, issued fewer commands per minute and missed more of the alarms than those using the function keyboard. But accuracy using the two devices was very similar. This is a credit to the robustness of the Votan VPC 2000 voice card, which uses a relatively unsophisticated recognition algorithm.

Discussion

Unfortunately, the study was somewhat limited in that a tone sounded to indicate that a word had been recognised. This very obtrusive form of feedback slowed ASR use down, so that it took approximately three times as long as function keyboard to enter commands (1.45s. vs. 0.5s.). This would explain the differences in performance of the input media; all performance measures used were time related. This is regrettable, but we were unable to device performance measures which could capture the same performance of the two media.

There were highly significant differences between performance of men and women using ASR. This fact has been noted by previous researchers (Talbot, 1987), and is presumed to reflect differences in the vocal tract lengths of male and female speakers. This, in turn, produces speech of different frequencies. It may be that ASR devices a 'tuned' to male speech, and cannot accomodate female speech. This is a problem inherent in using frequency based analysis of speech, e.g. in the use of filterbanks for analysis. While these sex differences are problematic, it is proposed that they are not researched further in this thesis. Developing technology may eliminate such differences, or it may be necessary to select personnel who prove to be efficient users of ASR. This is not a call for the use of ASR exclusively by male operators; there are instances in which some women will obviously achieve better performance scores than some men. Rather, it is a proposal that ASR be used by operators who can maintain acceptable performance with it.

Accuracy measures recorded prior to the experiment, and responses to the questionnaire suggested that subjects found ASR as easy to use as the keyboard, but more stressful. Zaijceck (1989) found that prospective users would prefer to use ASR away from other people, at least until they were confident of using it. This could be due to the public nature of speech. When typing, it is difficult for other people to assess one's performance without reading the screen; with ASR, the number repetitions one must make will indicate how well one is performing.

The overall performance scores indicated that the function keyboard was more suited than ASR in this context. The main reason for this finding was the large time difference between the two media. However, it can also be explained by the discussion above on speed of data entry. The task required users to select coded items with a colour and a number. This made the task very simple and more amenable to function keyboard activity, as would be expected from the results of Welch (1977).

A further finding from this study relates to one of the main arguments presented for the use of ASR; that it allows operators to perform 'eyes free' interaction with the device. In this study, the 'eyes free' component was actually found to be detrimental to performance. The function keyboard allowed subjects to structure their interaction far more effectively than did ASR.

The task required subjects to carry out a search of a simple, hierarchical plant diagram. While subjects using the function keyboard worked in a logical manner, those using ASR tended to skip between units and levels. The layout of the function keyboard led subjects to search the colour coded plant units in a consistent order. Therefore, in addition to providing a memory aid, a function keyboard can also help to structure behaviour by requiring a defined sequence of actions, e.g. always pressing 'red', 'blue' and 'green' keys in the same order. ASR has been proposed as a means of of focussing the operators attention away from the VDU. This study shows that this might actually reduce performance levels.

There are differences between ASR and Manual data entry on speed and accuracy measures. But these differences are strongly dependent on the two of task performed. This means that careful consideration should be given to defining and selecting tasks that ASR devices are to be used to perform, and to the interaction between operator behaviour and device performance. Additional consideration should be given to the possibility of using ASR in conjunction with manual devices. This could allow the operator to use each device to the best of its ability. It could also allow the performance of two tasks at once. In chapter six, we have seen how ASR can be profitably paired with physical tasks requiring the manipulation of objects. One could suggest that ASR could also be paired with more cognitive tasks.

Furthermore, ASR performance decreases when the operators' speech changes, for example, when they are under stress. Using ASR with other devices could allow them the choice of using a keyboard in emergency situations and ASR for routine operation. This raises two questions concerning the use of ASR with manual devices. Firstly, can ASR be used in conjunction with manual devices to perform dual task; secondly, could operators switch between ASR and manual use under different task conditions?

ASR use in conjunction with other devices

Damper et al. (1985) compared the use of a function keyboard with ASR to perform certain control functions, such as entering 'style' commands, and keyboard to enter the subtitles. They argued that ASR should be able to replace the function keyboard. Making dual operation possible; the subtitles could be typed, and style commands spoken. However, it was found that performance time actually increased by 9% when using ASR.

This increase in time was accounted for by two reasons: low confidence in using ASR led subjects to spend lengthy periods checking what had been recognised. The point relates more to the issue of adequate user training than ASR performance. Training is discussed in detail in chapter thirteen. The second explanation of their results presented by Damper et al. (1985) was that speech entry may impose additional cognitive demands on the operator than using a keyboard. This last point can be considerd in terms of the discussion of workload in chapter fourteen. The style commands were fed back to the subjects on a textual display, this required them to divide their attention between two screens. If they had used a single screen, then one could expect different results. This screen would display the main subtitle text, with style commands simply altering the appearance of the text. ASR feedback would, therefore, be integrated into the task of subtitle composition. Feedback is considered in depth on chapter eleven.

Several studies have investigated the potential of ASR as an adjunct to CAD systems. Dillon et al. (1990) devised a simple CAD type operation which required subjects to link points using lines of different colours. The lines were linked using a mouse to move a cursor on the screen. Colour selection was carried out by selecting options from a menu, using one of the following methods: the drawing mouse, a second mouse, a touch panel, or ASR. The use of ASR and touch produced the fastest and most accurate performance. As the touch panel required a great deal of screen space, Dillon et al. (1990) concluded that ASR was the best medium for the colour selection task.

Martin (1989) compared the use of a keyboard, mouse and ASR on a circuit diagram drawing task. She concluded that ASR allows the user to switch attention between the screen and reference material more easily than either of the other two options, and that users of ASR were able to complete more of the sets tasks under the time constraints of the experiment.

van Nes (1986) compared the use of speech and keyboard for script annotation. He found that speech allowed more annotations, and that the annotations were, on average, longer than typing. Further, ASR provided a faster means of annotating script, allowing subjects to enter complete sentences and to check their entry as it appeared. It is noteworthy that while typed annotations had a mean length of around half that of the spoken ones, they did not contain any less information. Thus, ASR produced more verbose but not necessarily more informative annotations.

Little and Cowan (1986) carried out a series of studies into the potential of ASR in helicopter cockpits. Although they concluded that the benefits of using ASR were not immediately apparent, they felt that this resulted form the fact that ASR was being used in conjunction with other devices. For instance, in high workload situations, Little and Cowan (1986) noted that,

> "some pilots tended to use their hand to select the display element and then perform the data input and execution command by voice, thus the keyboard search task was eliminated."

Thus, as well as being used as a primary method of data or command entry, ASR was used to perform verbal tasks in conjunction with manual selection tasks.

The difficulty of integrating voice with other media is confounded if one requires tasks to be performed simultaneously. The issue of dual tasks performance is addressed in chapter fourteen. It should be noted that the findings of Little and Cowan (1986) are based on the allocation of function between ASR and keyboard control on separate aspects of a task.

Speech can be used in conjunction with other input media, but care must be taken to allocate functions appropriately. It can be used to enter data or issue commands, but should not be employed to such an extent that vocal fatigue may ensue (Zarembo,1986). ASR should be used in such a way as to preserve the coherence of the task. It is pointless using ASR when other media can be use to perform the task as easily and more reliably.

When to use ASR

It has been argued that ASR can be advantageous in complex environments, because it helps distribute cognitive loading across separate processing modalities (Simpson et al. ,1985) . We have seen that ASR appears best suited to 'complex' verbal tasks, such as issuing commands. Not all 'complex' tasks benefit from the introduction of ASR, as study one demonstrated. Therefore, the dichotomy between simple and complex task proposed by Welch (1977) needs to be treated with some caution.

It is clear that verbal commands are not ideal for the manipulation of objects (Bierschwale et al. 1989). It is easier to move a cursor using a mouse or arrow keys, than to say "up. up. up. stop". This said, it is not immediately clear which tasks can be performed verbally and which manually. One option is to construct systems which rely on either manual control or ASR and see which gives best performance, but this will be costly and time consuming. It will be far more sensible to define design principles which can inform decisions concerning the use of ASR.

A series of studies has been carried by Wickens and his colleagues to suggest that there may be a psychological principle which can be used to assess whether or not an application is suitable for ASR use. This principle relates to the fact that performance is best when stimuli are paired with appropriate responses, in terms of human information processing codes. One can define stimuli as being either verbal or visuospatial.

Working memory theory (Baddeley and Hitch, 1974) proposes that there are two distinct codes of representation in the human information processing system (Verbal and Spatial). This leads to the hypothesis that spatial tasks will be best paired with manual responses, and verbal tasks with verbal responses. Chapter fourteen discusses this principle and related research in more detail. However, the S-C-R compatibility principle can be taken to support the assertion made in this chapter, that ASR ought to be most useful in situations involving 'complex' verbal tasks. By analysing the existing workplace, it is possible to define and describe tasks sufficiently to begin assessing potential applications for ASR. Chapters eight and nine present a study, based in an Electricity Grid Control Room, addressing the potential of ASR in a control room system.

CHAPTER EIGHT

A STUDY OF THE FEASIBILITY OF ASR IN GRID CONTROL ROOMS

This chapter reports study two, in which an assessment of possible application areas for ASR in grid control operation is undertaken. Grid control operators' tasks were assessed, using Hierarchical Task Analysis. The rationale behind this methodology, and reasons for using are discussed.

For the "dispatch operations", it was felt that the operators tasks could be made significantly easier with the use of ASR. This was particularly true for telecommand operation.

For the "loading operations", it was felt that although several tasks indicated possible areas for application, e.g. data entry and menu selection tasks, the introduction of ASR would not make a significant difference to performance of these tasks.

In conclusion, there are many possible areas in which it can be applied beneficially; telecommand is a particularly suitable application.

Introduction

Speech based interaction with computers is suitable for a number of applications (chapter six), and offers potential advantages over manual control (chapter seven). It seems useful to have a technology which allows the operator to interact with the main computer whether he is looking at the back panel, or at the screens on the control desk. The only technology which allows such mobility is ASR. In addition to allowing mobility, ASR can also allow operators to input data at the same time as they are reading them, thus removing any problems of trying to remember figures and the likelihood of reading errors.

ASR can be seen to have a number of potential benefits, but one needs to consider how well it compares with conventional input devices in terms of accuracy and ease of use. Study One reported in chapter seven shows that ASR is not necessarily suited to all types of process control tasks. It was concluded that ASR ought to be used for 'complex' verbal tasks.

There seems little value in using speech technology merely as a 'gimmick' in tasks which can be accomplished efficiently using existing technologies, or to force fit speech technology to existing systems, as Simpson et al (1985) point out,

> "The selection of potential task for speech recognition should be based on specific task requirements. Speech is not a useful substitute for manual data entry when such tasks are already being performed successfully... Speech input is likely to improve system throughput only in tasks that involve high cognitive, visual and manual loading."

Thus, what is required of a feasibility study is that it examines existing tasks, and the physical environment in which they occur, to determine which tasks will benefit from the introduction of ASR. The environment of grid control room appears to involve the characteristics which Simpson et al (1985) define, thus lending itself as a prime target for investigation. Study two will describe some of the tasks found in grid control operation, and the suitability of ASR for performing these tasks. It should be noted that the study is not intended to provide an exhaustive description of grid control activity, but acts as a method of filtering prospective areas for application.

It might be argued that the tasks of operators in grid control centres show certain similarities to those of inspection tasks (reported in chapter six). Industrial applications take advantage of the potential of ASR to offer operators a "third hand" (Nye, 1982). These potential advantages can be used to justify the use of ASR in situations involving medium to high physical workload, as it can offer an extra channel of communication between operator and computer. The capability for 'eyes free / hands free' data entry, offered by ASR, might be beneficial in grid control operation; data is gathered from several sources, and operators are required to move away from their work stations.

Industrial inspection tasks require the inspection of individual items, according to organised schedules. In other words, they consist of isolated operations on specific items. Grid control concerns a dynamic process. In other words, it concerns continuous operations on a constantly changing process. The majority of the tasks involved in grid control are planned before operation. Only around 10% of the activities involve unplanned actions (Sterling, 1978) and even these actions are usually carefully considered by the operators before being performed.

One can characterise inspection tasks as predominantly physical; items are manipulated and measured. But grid control is a collection of cognitive tasks. The physical demands made on the operator are minimal (with the exception of dressing switching diagrams). Rather, the operators' tasks involve activities such as planning, decision making, and monitoring. Therefore, physical criteria alone may not provide enough data for full consideration of the potential uses of ASR in control room systems. Nor is it sufficient to make applications decisions solely on personal experience. Some structured method of analysis is necessary in order to ensure that both cognitive and physical aspects of the domain receive sufficient consideration. While this might appear obvious, there is still much research required of human factors to investigate such topics. As Moray (1981a) points out,

> "...while new technology offers outstanding opportunities to help the human operator, almost nothing is known about how best to use the new displays and data entry devices. Only when appropriate new ergonomics have been done will the...design of systems rather than the design of components, be possible."

An initial familiarisation visit to the site used in this study suggested that there were some tasks which could provide suitable areas for the application of ASR. It was felt that further studies were required before any recommendations could be made concerning the potential use of ASR. Study two took the form of a Hierarchical Task Analysis. Four grid control operators were interviewed over a period of two days in February, 1989. They were asked to describe their work in terms of goals and the tasks which were required to achieve these goals. In addition observations of the operators' use of data sources and computer input devices was performed.

Hierarchical Task Analysis

For human factors, the analysis of behaviour at work takes the form of task analysis; a term which covers a multitude of techniques. Practitioners draw on varying combinations of these techniques, depending on the tasks and domain to be analysed. They can use direct observation or structured interviews to investigate operators' use of the present system. They can refer to operating data or critical incident reports to look at operating trends. These techniques can be combined to form the basic building blocks for system design, development, and evaluation.

There is little consensus among practitioners as to what constitutes task analysis (Shepherd, 1989). Arguments exist as to whether it should concentrate on gathering information about the tasks, or on ways of representing such information. Furthermore, there is controversy over whether task analysis should seek to describe actual behaviour, or possible behaviour, or means of achieving system goals.

We assume that an intrinsic goal of any form of task analysis is to capture 'real behaviour' of operators at work. This will allow decisions concerning system design to be informed by actual work practices as opposed to either assumed, or recommended practices. Operators generally develop their own methods of working, within the obvious constraints of existing systems and operational procedures, which means that the limitations of recommended work practices are overcome in a creative manner. Task analysis should attempt to capture this 'real' behaviour, rather than reflect written procedures.

Task checklists have been employed in several studies of ASR applications (e.g. EPRI report 1986). Such an approach can emphasise the physical aspects of the tasks in a given domain, concentrating on the manual aspects of data entry and operator mobility. They are either designed specifically for a domain, and so too specialised for general use, or are generalised, lacking the fine grain of analysis required to assess many situations. .

Fleishman and Quaintance (1984) recognise at least three definitions of the term "task". The first sees a task as a "set of conditions that elicit specific activities" from the operator. The analyst can assess how well the operator can be expected to perform these actions, in terms of general human abilities. Therefore, a technique which permits a general, well structured approach to all situations and tasks is required.

The second definition sees a task as a "process consisting of interrelated activities." The multidimensionality of activities involved in the performance of tasks can complicate behavioural analysis enormously, with problems facing the analyst concerning what aspects to emphasise or even include in the analysis. From this definition, analysis concentrates on the operators behaviour. The analyst can attempt to map perceived cues from the system and desired actions onto operator behaviour. Behaviours, which are important for the performance of specific tasks, can then be catalogued.

In order to structure the investigation of the various tasks required in grid control operation, a method of analysis known as hierarchical task analysis (HTA) was used. HTA was originally developed as a tool to aid the design of training routines, and has been extended to cover the tasks of operators in process control. As a training design aid HTA emphasises the information and activity requirements of personnel in specific environments, and allows training regimes to be designed which develop these requirements. In this study, HTA is used to outline the information requirements of operators (in quite broad terms), and relate these to the computer-based and communication activity in the control room.

If one is investigating the potential of a new input device, one needs to know what type of input activities it might be called upon to perform. Thus, initial investigation can attempt to map the new input device onto existing operations. However, speech input can offer a radically new method of interacting with computers. For this reason, a simple mapping of speech input to existing tasks might not show the full potential of speech. HTA allows tasks to be broken down in such a way that one can see the likely cognitive activities necessary to perform the tasks. It might be that speech can offer support to the operator performing such cognitive activities.

In HTA, each task is assumed to have associated with it a set of goals, and a set of actions necessary to achieve these goals. By observing the operators perform their tasks, and by interviewing them, one can begin to break down tasks from overall goals to component actions. Each goal breaks down into subgoals, i.e. to achieve a desired goal one must satisfy some other goals. The subgoals then break down into actions required to reach those goals. This classification produces a hierarchy of goals and actions. One can use this hierarchy as an overview of operations required in specific environments, and the information requirements of operators in these environments. The hierarchy can be presented either as a hierarchical diagram, showing the structural relations between tasks, or as a table. We use the tabular format purely because it is easier to fit onto A4 paper: it will convey the same information. The hierarchy provides a means of representing task analysis information, but not of describing or classifying it. The tables are accompanied by explanatory text.

A basic problem of task analysis lies in extending the analysis from a description of the tasks being performed to a full analysis of the underlying behaviours necessary to perform those tasks. It is clear that control room systems rely heavily on conceptual skills, rather than perceptual motor ones, and it is these which need to be investigated with respect to ASR. The third definition of 'task' offered by Fleishman and Quaintance (1984) defines it as a "conceptual construct" recognised by the operator. This raises the issue of cognitive, or information processing, aspects of operator performance. Task analysis methodologies were originally developed to investigate physical tasks and behaviours. There is, at present, little workable research into cognitive task analysis (see Wilson et al, 1988 for a review). However, from existing research and theories of cognitive ergonomics we can describe the requirements of a cognitive task analysis. Information processing tasks can be grouped together under three main headings of: visual perceptual; cognitive; and communication.

Visual perceptual tasks involve the detection or verification of signals, in terms of movement or change, and recognition and comparison of such signals. In conventional control rooms, a mimic diagram provides a schematic view of system operation. The 'eyes free' use of ASR could allow operators to scan this panel when necessary. However, one must consider the frequency of such scanning, and the reasons for scanning. The former can be estimated through direct observation. The latter requires a more detailed analysis of cognitive behaviour.

Cognitive tasks involve the coding and analysis of data in order to facilitate the prediction of state changes, or development of hypotheses concerning the likely results of such changes. These require the estimation and qualitative calculation of measured values from plant states which, in turn, involves monitoring states, remembering previous states to compare trends or detect emergencies. In order to evaluate the operators performance of these tasks, one must be able to measure and quantify it. This can be done using a form of task analysis which emphasises the cognitive components of the task.

One of the earliest techniques of cognitive task analysis, developed by Card et al (1983), is based on a simple model of human information processing. Tasks are defined in terms of goal structures, operations (elementary physical or motor actions), methods of accomplishing goals, and rules for the selection of a method. The technique is known by the acronym GOMS. The operator has a set of goals, often set out as operational procedures, and knows the actions necessary to achieve these goals. We suggest that the operator then formulates a plan to achieve the goal. This can be mapped onto the technique of HTA described earlier. Planning involves the analysis of the state of the whole plant, and in order to describe the behaviour needed to perform this activity, one must be able to classify knowledge requirements in the operators.

Rasmussen (1983) offers a very useful classification of cognitive behaviour which relates behaviour to the type of knowledge it requires (see also Crossman, 1956 discussed in chapter one). At the lowest level, is the automatic, skill based behaviour in which "highly integrated patterns of behaviour" are performed without conscious control. For example, an experienced driver changing gear. The next level is rule based behaviour, where a sequence of actions are grouped under a familiar rule for routine activities. Finally comes knowledge based behaviour in which a goal is explicitly formulated to cope with novel situations. Each level requires slightly more cognitive effort from the operator.

Hoc (1987) points out that the goals of operators in process control are often, of necessity, poorly defined. The system is too complex to be reduced to a simple set of rules and goals, and operator's areunable to see all the system actions. They need to rely on knowledge of the effects of their actions, knowledge of the system operation, and knowledge of the effect of existing controls.

In his study, Hoc (1987) carried out a detailed observation of operator behaviour, and then interviewed operators. From the interview, he was able to define the operators actions and classify knowledge into three types, input / output of data, states of physical components, and the evolution of system states. Relationships between these knowledge types was mainly causal and allowed operators to form a network for reasoning and planning.

Allengry (1987) also interviewed operators and analysed the transcripts to develop causal reasoning chains. From these he classified knowledge used in process control as relating to the actual process, basic principles of system functioning and the state of system components. These are very similar concepts to those put forward by Hoc. This gives an idea of the view that operator have of the system they are controlling and what information they will find useful.

Finally, information must be communicated to other operators and to the computer systems in use. EPRI (1986) saw this as a possible source of interference for ASR, especially if the operators are switching between telephone and ASR use. In situations where operators need to change their style of speaking, i.e. between phone and ASR use, it is not a good idea to employ ASR.

From the studies by Hoc and Allengry, and our own observations, we conclude that the majority of control room activities are of a supervisory and diagnostic nature. Incidents which lead to malfunction or breakdown can be kept to a minimum by carefully tracking plant behaviour and making appropriate adjustments to operations. However, even in well run control rooms, incidents occur. Such situations are stress inducing and call for quick, accurate diagnosis and command selection. It is known that stress can severely affect the use of current ASR devices (see chapter fifteen), but could ASR be viable, in principle, given adequate technology?

In certain situations, it is useful to be able to think aloud while solving a problem. However, talking aloud can often interfere with thought processes in other situations, e.g. trying to hold a conversation while driving in a strange town. The first question then concerns the relevance of speech to the problem being solved. The second concerns whether using ASR can interrupt problem solving. Until we can provide an answer to the question of the effect of ASR on problem solving, and until the ASR technology can be shown to handle the variations in human speech due to stress adequately, then one can rule out the use of ASR in "incidents". For this reason, applications must be selected from the many routine, supervisory tasks in control room systems, although the issue of ASR use in high workload situations in considered in chapter fourteen.

Overview of Grid Control Operation

A comprehensive definition of power system control is offered by Sterling (1978) thus,

> "Power system control is required to maintain a continuous balance between electrical generation and a varying load demand, while system frequency, voltage levels, and security are maintained. Further, it is desirable that the cost of such generation should be a minimum. The variable nature of the consumer power demand necessitates fluctuations in the total generation in order that the power balance be maintained."

The tasks of the operators in grid control come under the general heading of supervisory control and data acquisition (SCADA). This requires the collection of data from the plant operating in the area, e.g.in terms of status indications and measured values, and monitoring of system performance, e.g.trend and event monitoring. One can define the function of grid control as the safe, economic management of electricity power distribution, in terms of consumer demand and system constraints (Cegrell, 1986). The task of grid control is to track the variations in consumer demand as accurately as possible, and ensure that supply is matched to these variations. Generated electricity can not be stored in large quantities, so if too much power is generated the excess will be lost resulting in expensive waste.

Durley Park Grid Control Centre

From the above overview of grid control operations, it can be proposed that grid control comprises several tasks. These can be grouped under two main headings: loading and dispatch. This grouping is reflected in the layout of the grid control centre at Durley Park. The control room is divided in half: with loading operations on one side of the room and dispatch operations on the other.

The tasks are not mutually exclusive and communication between operators is often required. Further, operators rotate on a fortnightly basis between dispatch and loading. The control room is under the control of a supervisor who is assisted by a shift clerk, two loading operators, and two dispatch operators. This study concentrates on the activity of the loading and switching operators The role of the shift clerks is to compile meteorological data which could effect electricity distribution. Their work is based on telephone conversations with weather stations, and a computer link with National Control. Their use of the computer was too limited to consider it in terms of input technologies. The supervisors' role is mainly administrative, which again required minimal computer use.

Study Two

This study was performed with the assistance of grid control staff at Durley Park. It represents a summary of routine work in the grid control room. HTA provides a 'snapshot' of activities, and the discussion in this study is not intended to capture the full detail of every operation and activity. HTA is used to break the different activities into component tasks. These tasks can then be assessed in terms of their viability for ASR, using the concepts and criteria developed in chapters six and seven.

Loading Operations

Loading is primarily a planning task. Planning begins five years before the operation date. The likely consumer demand is estimated based upon demand on previous days with similar conditions, i.e. weather, time of year etc. These plans are revised yearly until six weeks before the operation date.

From six weeks, plans are revised weekly in line with the current demand trends, long range weather forecasts, plant commitment etc. These plans are drawn up in the centre's planning department. A set of weekly notes is delivered into the control room on a Friday afternoon. This represents the culmination of the work of the planning department. Loading operators have to put these plans into operation, and make decisions concerning any deviation from the plans, e.g. if cloud cover increases demand will rise with lighting requirements.
The table of Loading Operation which resulted from the HTA is shown below.

Superordinate	Operations, defined by HTA	Notes
0.	**Loading Operations**	
	1.Operational Planning	
	2. Operational Control	
1.	**Operational Planning**	
	1. Monitor Station Availability	
	2. Calculate Demand Estimates	
	3. Manage Power Requirements	
2.	**Operational Control**	
	1.Monitor System Performance	
	2. Effect Changes in System Performance	
	3. Arrange Import or Export of Power	
	4. Maintain Transfer Value	
	5. Ensure Economic Distribution	
1.1.	**Monitor Station Availability**	
	1. Receive Data from Stations	
	2. Receive Telephone Calls from Stations	
1.2.	**Calculate Demand Estimates**	
	1. Modify Historic Curve	
	2. Run GOAL	
1.3.	**Manage Power Requirements**	
	1. Run GOAL	see 1.1.
	2. Receive Data from National Control	
	3. Receive Weekly Notes	
2.1.	**Monitor System Performance**	
	1. Receive System Frequency Readings	Freq. >50+/-2 Hz see 2.2
	2. Run Simulations of "credible faults" on GOAL	
	3. Monitor System Security	Switching 4
2.2.	**Effect Changes in System Performance**	
	1. Telephone Power Stations	
	2. Receive Telephone Calls from National Control	

	3. Complete Log	
	4. Monitor Station Availability	see 1.1.
2.3.	**Arrange Import / Export of Power**	
	1. Monitor Transfer Values	
	2. Calculate cost in line with Merit Order	
	3. If cheaper to 'buy' in than produce power, then arrange import	
	4. Arrange Export, if required	
2.4.	**Maintain Transfer Value**	
	1. Receive Transfer Value from National Control	
	2. Receive Data from 'transfer display'	
	3. Complete Log	
2.5.	**Ensure Economic Distribution**	
	1. Receive Merit Order from National Control	Aim to maintain or better, position in Merit Order
1.2.1.	**Modify Historic Curve**	
	1. Calculate Probable Demand Peaks	
	2. Map Probable Demand Peaks onto Previous, Similar Day	
1.2.2.	**Run GOAL**	
	1. Receive Demand Estimates	
	2. Update GOAL	Input data at set times
	3. Amend GOAL	Data from Stations, via GI74 or telephone

At Durley Park Grid Control Centre, loading operations are shared between a Dispatch Operator and a GOAL operator. The discussion which follow develops the HTA presented above, in terms of this distinction.

Dispatch Operation

The tasks of the dispatch operator are mainly concerned with energy disposition. The dispatch operator calculates the demand estimates for the peaks and troughs throughout the day. Operators are given 'transfer values', calculated to match supply to demand, from national control which represent targets to be met every half hour.

In order to meet the demand as economically as possible, it is necessary to schedule the starting up and shutting down of plant as accurately as possible. Demand for electricity varies with time of year (i.e. increased heating and lighting requirements in Winter), and time of day (i.e. peaks of demand at breakfast time and at nightfall). These variations provide the basis for the planning of likely consumer demand. National control calculates a weekly 'merit order' which lists available plant in terms of economic and efficient production of power.

However, it is not always possible to follow the 'merit order', so the dispatch operators need to ensure that the cheapest alternatives are considered and used. The dispatch operator checks the running and availability of power stations. They monitor the system frequency, which serves as an indication of any imbalance between demand and generation, and needs to be kept around 50 Hz.

They organise exchange of power with other areas. If too much power is generated in an area, i.e. if demand drops too quickly for the gird control centre to follow, operators must arrange for power to be exported to other areas so as not to waste it. If not enough power is being generated or if it is cheaper to buy in power from another region than to use expensive plant then power may be imported to the area.

i.) GOAL Operation

GOAL (Generator Operation And Loading) is a nationally run computer program which models consumer demand throughout the national grid during the day. It calculates the most secure, economic combination of plant per area. In order for its calculations to be accurate it needs to be updated at regular intervals during the day, i.e.4am, 9am, Noon, 6.30pm. The data necessary to update GOAL are collated by the GOAL operator, and include the demand estimates calculated by the dispatch operator, and any variation in measured values from the power stations.

In addition to updating GOAL, the operator manages power station performance in line with GOAL data. The power output levels from the stations are fed directly to the control room via the GI74 network. The operator has the area's generation summary displayed on the GI74 computer. If changes are spotted the operator telephones the station to ascertain where the changes have occurred and what action is being taken for them. If a fault occurs, the operator feeds this information into GOAL. Commands issued by the operator and any changes in the system are logged by the GOAL operator.

ii.) Switching Operations

The two operators on the switching side do not share separate aspects of the same task, as do loading operators. Rather, the operators carry out the same tasks, although obviously not on the same equipment. There tasks can be summarised under three headings: ensuring system security; ensuring that voltage production meets requirements; and managing outages. The results of the HTA for switching operations is shown below.

Superordinate	Operations defined by HTA	Notes
0.	**Switching Operations**	
	1. Arrange Outages	
	2. Manage Outages	
	3. Monitor Voltage Production	
	4. Ensure System Security	
1.	**Arrange Outages**	
	1. Receive Weekly Outage Dairy	
	2. Telephone National Control for permission	
	3. Telephone manned substations	

2.	**Manage Outages**	
	1. Telephone National Control for permission	
	2. Telephone manned substations	
	3. Issue Telecommands	
	4. Log Commands	
	5. Issue Commands	
	6. Dress Switching Diagram	
3.	**Monitor Voltage Production**	
	1. Receive Weekly Report	
	2. Receive Grid Control Area Summary	via GI74
	3. Monitor for Alarms	on GI74
	4. Monitor Line Loadings	on GI74
4.	**Ensure System Security**	
	1. Complete RISSP reports of activity	
2.2.	**Telephone Substations re. Outages**	
	1. Log Outage Commands	
	2. Issue Outage Commands	
	3. Check Log with Substation Staff	
2.3.	**Telecommand**	
	1. Log Telecommand	
	2. Select unit on GI74	search through menus
	3. Type Telecommand	
	4. "Double Send"	
2.6.	**Dress Switching Diagram**	
	1. Climb Ladder to Unit	
	2. Select Switch	
	3. Change Switch Position	
	4. Label Changed Switch	

Ensuring system security can be seen as the overall goal of the switching operator. If any fault occurs, it is their job to manage the change in supply and the introduction of new plant. The voltage profile provided by the GI74 provides enough information for changes to be spotted. Monitoring the voltage profile of the system is the default activity of the loading operator.

Outages are prearranged times when plant are removed from the grid for maintenance or construction. It is the duty of switching operators to manage such outages. The outage diary arrives on the Friday before operation date, and a daily outage sheet updates this data. The outages are not timed. The operators wait for telephone calls from personnel on site to request permission to begin outages. If plant is part of the supergrid, then the operator needs to get permission from national control.

Outages require switching between plant and between bus bars, and the isolation of various parts of plant. Some switching is done by the personnel at the plant, and is coordinated over the telephone from the control room. Other switching can be carried out directly from the control room using telecommand. Because switching operators assume responsibility for the switching they control, an accurate log of their commands needs to be kept. For this reason the operators use a command language which requires a simple syntax and restricted vocabulary to avoid any ambiguity. Instructions are given over the telephone and written in a log using this command language.

Scope for ASR in Grid Control Operation

i.) Loading Operation

The majority of the dispatch operators' tasks are carried out using paper and telephone. They use a VDU to monitor system frequency and a summary of generation over the whole area, but do not need to input data to the computer. Telephone conversations concern the status and availability of plant or enquiries to national control. Data arrives on computer printout (weekly notes from planning department, merit order from national control), or teletype printouts (transfer values from national control). Operators enter data into a dispatch log and provide the GOAL operator with demand estimates.

It was felt that the minimal computer interaction involved in the dispatch operators work did not offer much scope for study of input devices. Whilst logging can be time consuming, and may be prone to errors in periods of high workload, it is not felt that using ASR to log data will significantly effect dispatch operation performance. There are ASR devices on the market which allow 'hands free' dialling of telephone numbers. The switchboard used to connect operators to power stations could be replaced by a voice dialling system. This could save the operator some time, but time is not often seen to be a crucial variable in the control room. Thus, voice dialling could be viable but is not necessarily desirable.

ii.) GOAL Operation

The operator selects an option from a numbered menu. Function keys allow the operator to return to local or function menus. Data entry is carried out by editing existing displays. The cursor is shifted line by line to the appropriate place, and data typed over existing data. Although this seems rather a clumsy method of text editing, the overall, highly constrained interaction does not present the operator with too many problems. ASR could allow the operator to directly input demand estimates and could provide a facility for logging actions. However, these suggestions will not significantly improve task performance.

iii.) Switching Operations

There are several aspects of switching operation that could benefit significantly from the introduction of ASR. The computer system used to operate telecommand is also used to call up displays on GI74. The system has an extremely long system response time which slows down work in time of high workload, and the procedure for calling up pages of information is also time consuming.

If ASR was used to execute telecommands, the commands could be issued and logged simultaneously. In periods of high activity, it is easy to overlook the fact that all commands must be logged. The command language already in use would be highly suitable for use with ASR devices. Further, this system could be used to display the data the switching operators require. The development and assessment of this system is described in chapter nine.

Switching accumulates at certain times during the day, specifically at the start and end of the working day . This often leads to unnecessary waiting time for the control room operators. Part of the reason why switching is time consuming is that it requires permission form other sources. If permission was sought during quiet periods, as part of the planning phase of switching, time could be saved.

The system which the control room is responsible for is displayed on a mimic diagram on the wall in front of the switching operators' desk. This diagram shows the plant on each of the different voltages lines and the switches and isolators in these plant. It is used to indicate the current status of the system. If any changes are made to the system, i.e. if switching operations are performed, the diagram must be altered, or dressed. This involves the operators climbing a ladder to the point at which the change needs to be made. Dressing the back panel could be carried out from the control desk either using a keyboard or voice control. This would remove the need for operators to climb ladders to dress the diagram. However, it might be that the physical act of climbing a ladder and altering the diagram provides the operator with a sense of task completion, i.e. the task is felt to be completed when a physical action is performed. If the task is to be carried out from the control desk, an additional consideration would be suitable termination cues.

This physical termination activity is also found in the operation of telecommands. Here the telecommand to be sent is set up on the computer, and the operator is required to press "send" keys on two separate keyboards before the command will be sent. This is intended as a safety precaution to prevent errorful commands being transmitted. In the demonstration system using ASR, the operator is required to hit a touch pad on the screen in order to send the command. This would provide the operator with as much system security as currently available.

Conclusions

A large amount of operator time is spent in communicating with colleagues and with national control or power stations, over the telephone. Study two did not record the actual proportion of time spent in communication, but operators agreed with the figures proposed in the EPRI (1986) report. Person to person communication, i.e. talking to colleagues in the control room, takes around twenty minutes per hour, while telephone communication takes up between twenty to thirty five minutes per hour. This shows that around fifty percent of the operators time is already taken up with speech communication. As we saw from the study of police helicopter pilots (Linde and Shively, 1988), reported in chapter six, ASR will not be effective when competing with other speech requirements. Of the remaining time, operators are using the computer systems or dealing with hand written data. These tasks vary in relation to the role of the operator.

The division between GOAL and dispatch operation is somewhat artificial. In reality they represent two aspects of the same task, as evidenced by the high degree of communication between the two operators. The twin tasks of GOAL and dispatch operation which make up loading do not immediately suggest likely areas of application for ASR. However, elements of the task show the possibility of ASR use. Logging could be carried out using ASR, as could the direct entry of demand estimates to the GOAL program. However, the task of logging as it exists at the moment is not felt to be too onerous, and the data entry of GOAL could be improved without the use of ASR.

GOAL operation is highly structured and uses a limited vocabulary. ASR could very easily replace the keyboard in this task. Logging of commands could be carried consecutively with data entry, removing the need for paper records. But it is not clear that ASR will improve performance of the GOAL operations. Loading represents a set of cognitive activities. Operators are required to plan, monitor, and calculate generation and demand of electrical power. Operators stressed the need for thinking time to check their calculations and evaluate plans before execution. What would be required of any computer based device, if it was to make a significant improvement to operator performance, is that it supports these routine activities. Switching can be characterised as having more physical components than loading. Operators scan the back panel for information, and need to dress it to display changes. This requires 'eyes free' activity and operator mobility. ASR has much potential in switching activities in grid control. The fact that switching can be scheduled also means that there is little possibility of competing communications tasks. It offers a reduction in physical workload, and can offer a cognitive aid in automatic logging. Study three, reported in chapter nine, will address the viability of speech controlled telecommand operation.

CHAPTER NINE

THE DEVELOPMENT AND ASSESSMENT OF A SPEECH BASED TELECOMMAND DEMONSTRATION

A demonstration using ASR to control a simple telecommand scenario was designed; it was assumed that operators would find difficulty in assessing the utility of a novel technology without first having some experience of it.

Study three describes the results of a field trial, which consisted of a short demonstration by an experienced user of ASR followed by a performance test and an assessment sheet. The assessment was completed by fourteen operators.

The results show that operators could very quickly achieve acceptable performance. This ease of use was reflected in their responses to the assessment questionnaires.

The speech based telecommand demonstration shows that ASR can be used to restructure current tasks. Such restructuring can lead to improved performance by reducing the effects of computer use as a burden to carrying out primary tasks.

Introduction

From Study two (chapter eight), we can conclude that the majority of grid control room activities are of a supervisory and diagnostic nature; applications must be selected from the many routine, supervisory tasks in control room systems. Telecommand represents such a routine activity. Furthermore, telecommand can be described as a verbal task of some complexity. Operators must formulate and issue commands, according to a defined syntax, to effect changes in system operation. In chapter six, it was proposed that prospective ASR applications should consist of 'complex', verbal tasks. From this definition, telecommand ought to be ideal.

After selecting possible tasks, simulations using ASR need to be carried out to assess their viability.There are two approaches to evaluating the use of ASR in actual tasks. The first, and probably most common, is known as 'the Wizard of Oz' technique (see chapter ten). Here, the experimenter imitates a speech recognition system between subject and computer. Such studies can be useful in deciding the style of vocabulary necessary for ASR use, or in studies which could be adversely affected by the tendency of ASR devices to introduce errors. But the approach is limited in that it relies on a perfect ASR system, the human listener. Even if the experimenter tries to introduce errors, there is some difficulty in establishing a 'real life' interaction with an ASR device. The second approach is harder to set up, but ought to be recommended as allowing more realistic situations to be examined. Here, the subject controls a simulation of a process task using an ASR device.

Mitchell and Forren (1987), and Forren and Mitchell (1987) compared an ASR device with manual control in the simple task of page selection in a control room demonstration system. They found ASR to be much slower than manual control. This led them to conclude that ASR was inferior to manual control. However, one could offer several criticisms of their study.

The first being that feedback was provided to the operators on a word by word basis. This will inevitably slow down performance. If feedback was delayed until a whole command had been entered then performance would be faster (see chapter eleven). The issue of interaction time, while stressed by Mitchell and Forren (1987), may not be as important as they think. Our interviews with control room operators reveal that they often rely on "thinking time" to check the commands they have issued or are about to send.

One point which is clear from the Mitchell and Forren (1987) study is that, due to the occurrence of errors, ASR will require the operator to view the screen to detect and correct errors. It seems a little premature to rule the use of ASR because it fails to satisfy a physical criteria which may not be valid in control room systems. Most routine data entry will be carried out at the operators workstation, and thus, the issue of 'eyes free' use does not seem important. Primarily, Study Three focuses on the acceptability of an ASR system in terms of ease of use and length of training required, but consideration is given to important human factors in the design of ASR systems for operational control.

Telecommand is the activity of remotely changing the status of switches in substations from the grid control room. Operators can also issue commands to staff on site to perform manual changes. Both forms of remote switching are carried out using a standard vocabulary and syntax. The simple nature of the switching task, the standard vocabulary and constrained syntax, all serve to suggest that telecommand offers much scope for the use of ASR.

Overview of Demonstration

The telecommand demonstration was designed to illustrate how ASR could be used to restructure routine activities, making them easier and less time consuming than current practice. As was noted in chapter eight, the computer system operators currently use at Durley Park has a long system response time. Its use of a hierarchy of menus means that operators spend time scrolling through pages of irrelevant information rather than calling the information required directly onto the screen. In addition to computer operation, telecommands must be logged by hand. This can also be time consuming. In periods of high workload, operators can forget to log commands. This would suggest that some form of automatic logging would be desirable. We wanted to keep development time to a minimum for a number of reasons. It was decided to use a small, fictitious grid (as shown in figure 9.1), for the demonstration . The main components of the task of issuing telecommands were used as the basis for the demonstration.

The names of the substations are all place names in Birmingham. Each switch has an identification number. When the switch is open, it shows as an empty box. When it is closed, the switch becomes green. Switch status is also indicated by the position of the switch, and the operation of changing the status of the switch causes the switch to move from its current position to the one selected. Current practice requires operators to move down through the menus until they get to the switch status page. The use of graphic representation in this demonstration is easier to use and status information can be obtained directly. Admittedly, the small grid makes such a scheme relatively simple to use, but the display principle could be extended to a larger grid, with small sections being shown on the screen at a time.

The window above the grid control diagram contains the words issued for each telecommand. Once an acceptable command has been completed, a 'send' button appears in the bottom right of the screen, together with a message to allow the operator to check that the command being sent is the desired one. The operators touches this button to send the message to a printer where it is logged. Thus the additional task of logging commands manually is removed.

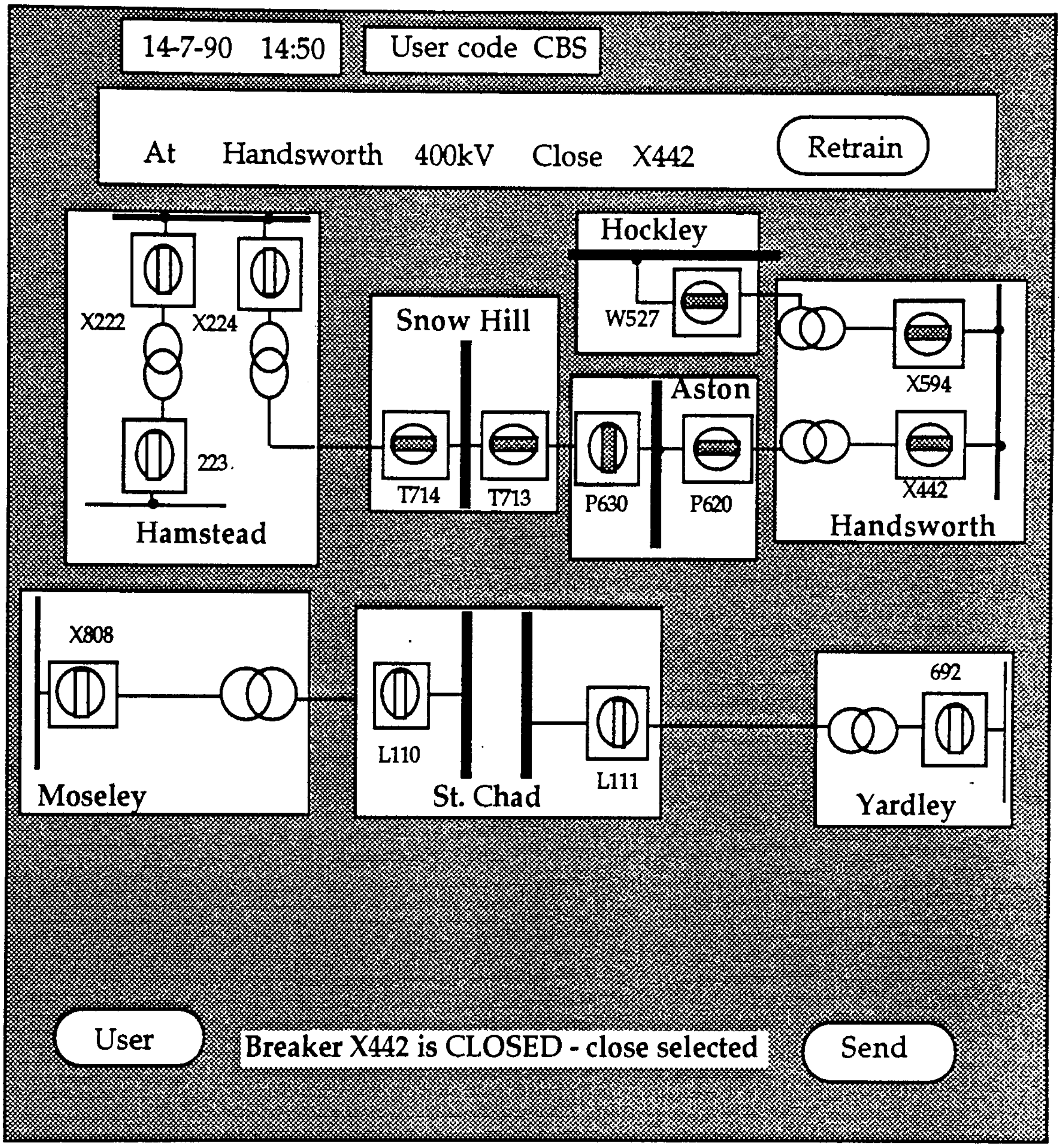
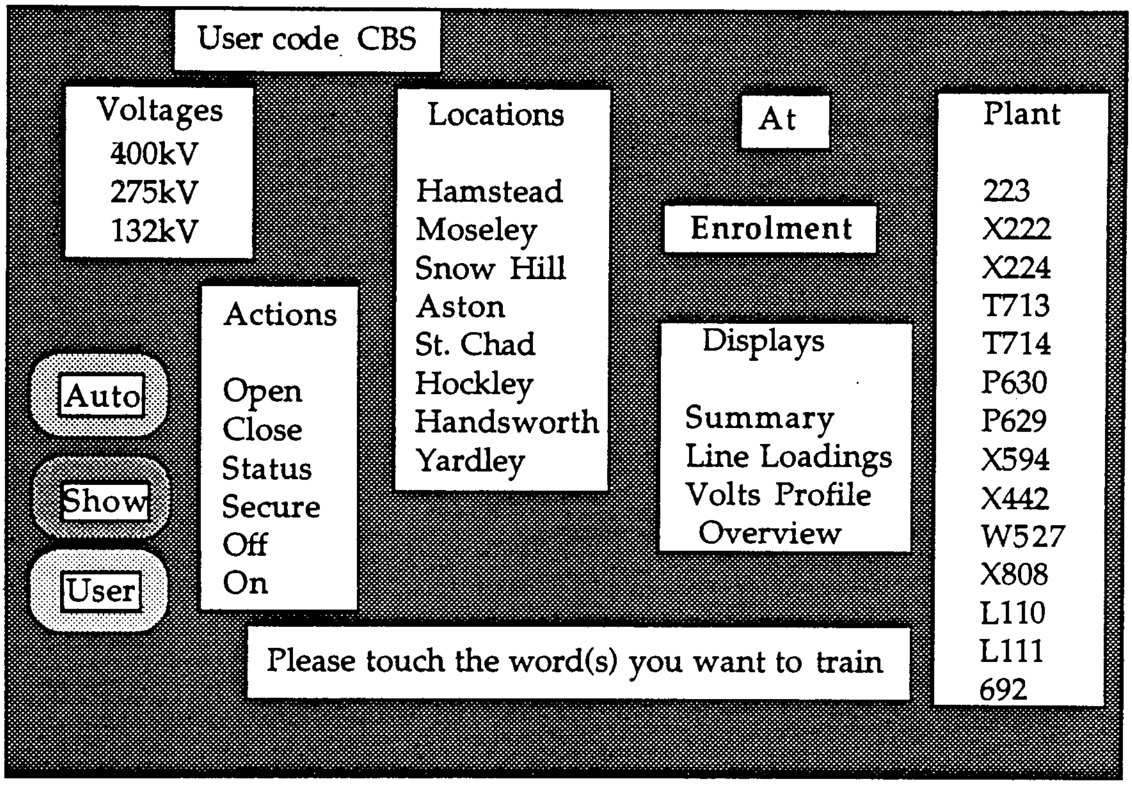


Figure 9.1: Schematic of Telecommand screen layout
[adapted from colour plate in Usher and Baber, 1989]

In addition to allowing operators to issue telecommands, this demonstration also gives information concerning line loadings and voltage profiles, together with an area generation summary. Pages of information can be called up using short, simple commands; this was the task investigated by Mitchell and Forren (1987). Therefore, as well as illustrating the use of ASR in a specific application domain, we can also show ASR to be a means of data entry or page selection, and a medium for command and control.

i.) Enrolment

The device used in this study is speaker dependent. This means that before using it, operators must first train it to recognise their utterances of the words in the vocabulary. This process is known as enrolment.



Please touch the word(s) you want to train

Figure 9.2: Diagram of Enrolment Screen [from Usher and Baber, 1989]

Prior to enrolment, operators create a personal identification code which is used to label their template file. The first screen of the demonstration contains a five by five matrix of touch buttons. Operators select three letter codes by touching the letters. As a letter is touched, it appears in a box at the top of the screen. When the operators touch this box, they are taken to the enrolment screen. The demonstration uses an automatic training schedule to reduce the amount of time operators need to spend in enrolment. Figure 9.2 shows the enrolment screen, with the words in the vocabulary in separate sections.

The operator touches the 'auto' button, and is prompted to say the first word, in the "voltages" column. Once the operator says this word, he is prompted to say the next word. This process continues until all thirty six words in the vocabulary have been enrolled. It is customary to repeat the process of enrolment several times in order to attempt to capture some of the inherent variation of speech. However, in order to keep the time operators spent using the equipment to a minimum, we only used a single pass enrolment. This means that if more time was spent enrolling the device, performance ought to be better than that obtained.

If a word had an error associated with its enrolment, the operator is advised to proceed with the enrolment process. Errors can result from the operator mispronouncing the word, coughing etc., or if the utterance exceeds the two second time limit built into the device. At the end of the enrolment process, words which have been correctly enrolled appear in red: those with errors associated with them appear in white. By touching the white words, the operator is able to enrol them correctly. Once the words have been enrolled, a 'telecmd' button appears on the screen (in place of the 'auto' button). Touching this button takes the operator to the telecommand demonstration. It is also possible to escape from the telecommand demonstration to re enrol problematic words. The operator touches the word he wishes to re enrol and speaks it when prompted. The 'telecommand' button then takes him back to the demonstration.

ii.) Vocabulary and Syntax

The device used in this study relies on pattern matching principles (see chapter four). The spoken utterance must be sufficiently close to a recorded template for recognition to occur. This could mean that similar sounding words, e.g. "one", "done", are potentially confusable. Notice that the acoustic similarity can be quite broad; it is not uncommon to find the words "four" and "five" confused.

Consequently, an ideal system would use a highly distinguishable vocabulary. However, in an application using a predefined vocabulary, e.g. place names and switch numbers, it is not always possible to redesign the vocabulary for optimum performance. In this demonstration the names of the substations are all place names in Birmingham. We decided to use real names rather than design artificial ones. While the latter would allow for acoustically distinct words to be used, providing maximum efficiency for the recogniser, it would not reflect real life performance. Real place names offer the possibility of errors and confusions in recognition, which are likely to occur in normal use. This would require some means of correcting errors, and is discussed below.

During the task analysis sessions, we noticed that telecommands were logged using a standard format. This format had been developed to reduce ambiguity in the recorded logs and spoken conversation, over the telephone, with station staff. It was decided to retain the standard format because the operators were familiar with it. The standard format of telecommands is as follows:

"AT <substation name> <voltage level> <perform action> <on switch number>"

The system has a vocabulary of thirty six words. The vocabulary was not designed to have minimal confusability, but contained some confusions possible in real grid control operation, e.g. "Handsworth" for "Hamstead", "Z111" for "Z110". The experimental task used 28 words: "open", "close", nine substation names and fourteen switch numbers.

iii.) Error Correction

As the vocabulary was known to contain possible confusions, we designed two methods of error correction which were simultaneously active throughout the task. Subjects could either repeat a misrecognised word or the whole command phrase. Further, a 'retrain' facility allowed subjects to escape from the task and update templates of problematic words. The desired word can then be selected and retrained. The user can then return to the telecommand page to resume issuing commands (the issue of error correction is discussed further in chapter twelve).

iv.) Feedback

Two types of feedback are used to inform the operator of the devices recognition performance. The words recognised are displayed in the text window at the top of the screen. They do not necessarily appear at the same time as they are spoken because the ASR device uses a memory buffer to store speech before analysis. This imposes a short time delay between words being spoken and displayed on the screen. The words appear in predefined order, according to the standard format, but can be spoken in any order.

In addition to displaying the words recognised, feedback also gives an indication of how well the words have been recognised. Recognition occurs when a spoken word is sufficiently similar to its stored version. Effective performance results when operators speak in a consistent manner. If words are well recognised the appear in white. If they are poorly recognised, they appear in grey. From observing operators use this equipment, feedback tends to be important primarily as an indication that the device is correctly recognising commands, rather than performing the actions. In other words, feedback is used primarily as a verbal signal. That is, providing the desired word appears, it does not matter how well it has been recognised. Operators can assume that once a command has been correctly recognised, the device will perform the desired action, i.e. close a particular switch.

The use of a 'mimic diagram' informs users of the status of the switches in the grid. This provided a form of plant feedback, concerning the current state of the grid (the many issues surrounding the appropriate use of feedback for ASR in control room systems are dealt with in chapter eleven).

Study Three

The equipment was taken to Durley Park grid control centre and set up permanently for seven days. The equipment was left for four days for the operators to investigate on their own, and three days were spent carrying out the assessment. Fourteen operators participated in the half hour studies.

Study ten, reported in chapter thirteen, found that the most efficient method of teaching people to use the equipment was for an experienced user to give a demonstration, keeping explanations of its working to a minimum. In the demonstration, the experimenter worked through pages of information and changed the status of five switches. Any errors that occurred were corrected, and one word was re enrolled. Following this, the operator enrolled the device, and completed a practice session involving the change in status of five switches. This allowed him to get used to speaking to the device, correcting errors, and re enrolling words.

Finally, a performance test was set up. This test was very simple and was devised so that each operator performed the same task. The operator was required to change the status of all fourteen switches in a set order It was not supposed to bear any resemblance to actual telecommand tasks which involve an element of problem solving, but aimed to illustrate how ASR could improve the sending of telecommands. Because accuracy was considered to be more important than speed, our main performance measure relates to how many words it took to issue a complete command.

As we saw in chapter five, one of the problems with carrying out experiments into ASR is the difficulty of defining adequate measures of performance. Most workers report the percentage of words recognised by the device (recognition accuracy), or distance scores between template and speech for the device used. But these measure can only be related to the performance of the ASR device, not the system as a whole. Also, because the distance score is calculated as part of the matching process, it is very difficult to know exactly what the score represents, i.e. does it give a score for the whole word or just prominent features, does the score provide a consistent indication of performance or just a 'one off' measure?

In this study, we decided to use a performance measure related to the actual task. In order to issue a telecommand, subjects need to speak five words, i.e. "At Moseley 400kV. close P808". If the device recognises all these words first time, then the command will be issued in five words. However, it is not possible to achieve perfect performance using any ASR device. We include a means of error correction which allows subjects to repeat individual words. Thus, if a word is misrecognised, the subject repeats it to complete the command, and the number of words used increases from five to six. We define acceptable performance limits as subjects not having to repeat more than two words per command.

i.) Performance results

Figure 9.3. shows the percentage of commands issued using various numbers of words, e.g. 51% of all commands were issued using five words. Our performance measure concerns the percentage of commands issued using seven words or less.

From the figure 9.3 it is clear that 72% of all commands issued fall within this criteria. In other words, operators only had to repeat two words in order to issue a complete command. One might ask whether the repetition of any command words is desirable, and whether one ought to strive for a system which achieves perfect performance at all times. Bearing in mind the complexity of recognising speech, the age of the equipment used, and the fact that we only employed a single pass enrolment, these results are very favourable. Possibly, the results would be better if we had used more modern equipment and more enrolments, but they would not have achieved 100%.

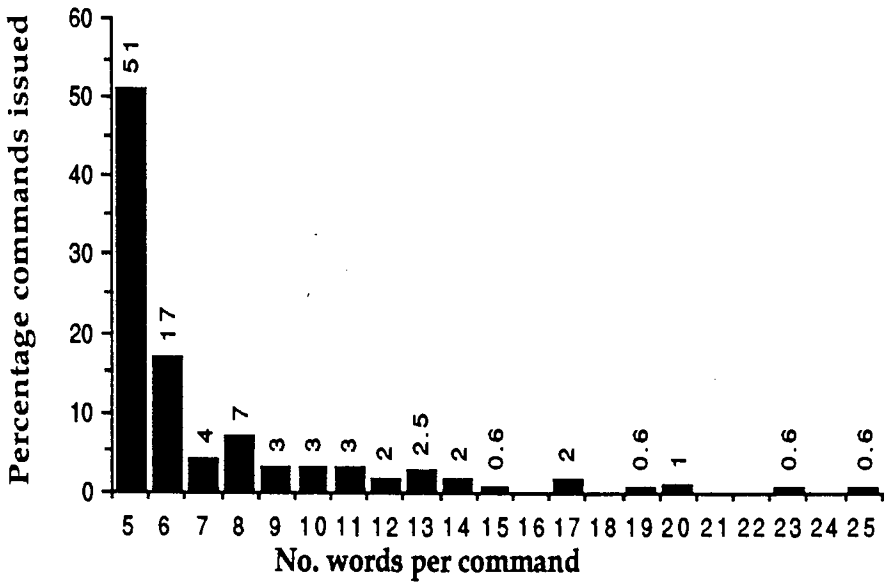


Figure 9.3 : Percentage of commands issued against number of words in each command.

Two assessment sheets were given to the operators after they had completed the performance test. The first involved a simple binary choice of words to complete a statement. The second required operators to indicate on a 50 mm. line their response to a statement. Results from all fourteen operators were pooled and the following conclusions drawn.

Operators were prepared to accept the limitations of the device. They found that it was easier to use and performed far better than they had expected. The results from the assessment questionnaires support these statement, with operators agreeing with the statement that one did not have to be highly trained to use the system. Operators agreed that ASR would make telecommand easier than at present, and could be used in other areas of grid control. Finally, they agreed that they would use the system on a regular basis if it was available.

Conclusions

The results from the performance test show that even after a short period of familiarisation, operators are able to use ASR to an acceptable standard of performance. No doubt such ease of use effected their responses to the assessment questionnaires, which were very favourable. If ASR was to be used in grid control rooms, certain considerations need to be given as to its implementation. The demonstration used a head mounted microphone on a lead to the ASR device. In the control room, operators are required to move away from their desk to the mimic diagram and the the desks of other operators.

To assist this mobility, one could use a radio microphone to issue commands to the device. A belt mounted radio pack would contain a switch to control the microphone, and a button to send the commands. This mobility would mean that operators would not always be able to see the feedback on their VDU. However, as one of the operators suggested, it would be possible to have a large, LED display mounted above the mimic diagram to provide verbal feedback concerning the commands to be sent. The issue of what constitutes appropriate feedback to people using ASR is dealt with in Studies Seven and Eight (chapter eleven).

A recent study by Usher (1990) found that the use of a radio microphone can degrade recognition performance. Alternatively, one could require telecommands to be sent only when operators are at their desk (which is currently the case). A desk mounted microphone could be used, to remove the trouble of putting on a head mounted microphone each time one wished to use the system.

It is obvious that our speech can change when we are under stress. This would result in our voice changing from that recorded in the templates. What is not obvious is how our voice changes; stress effects people in different ways (see chapter fifteen). It is difficult to compensate for stress in the use of ASR. For this reason, ASR should be used in conjunction with existing technology in order to allow operators to use proven devices during emergency situations. Such use has already shown some success (see chapter six).

In emergency situations we tend to revert to well learnt behaviour patterns: if these are dependent on existing devices then one would require to keep them until operators have learnt to deal with emergency situations using ASR. This would also be a useful testing procedure for ASR - if it was better than existing methods, operators would prefer to use, and the existing devices would become redundant. Unfortunately, such evolutionary device comparison does not seem practical in real working situations.

The response from the operators, and the results of the performance test combine to highlight the potential of ASR in control room situations. The speech based telecommand demonstration illustrates how ASR can be used to restructure an existing task in such a way as to make it easier to perform and less time consuming. The demonstration automatically logged commands, and allowed information to be accessed faster than the current system.

Issuing telecommand is not only a process of sending a command, but also a problem solving exercise. In a sense, removing the burden of logging commands can be seen as a simple cognitive aid. However, there has been very little work into the potential of ASR as an aid in cognitive tasks. Operators believed that ASR could be used in other areas of control room operation. Many of these tasks were discarded in study two (chapter eight), because it was felt that ASR use would not make a worthwhile improvement to existing practice. However, the use of ASR for tasks surrounding telecommand show its potential.

CHAPTER TEN

SPEECH BASED INTERACTION WITH COMPUTERS IN THE CONTROL ROOM

ASR has often presented as a 'natural' medium for HCI, because it employs speech. This argument is based on an erroneous view of both human use of speech and human computer dialogue.This chapter begins by offering a definition of "dialogue" in the context of HCI. It is that argued human computer dialogues are markedly different from those between people. This is supported by study four. The main characteristics of dialogues for speech based interaction with machines are outlined in study four. Study five considers the effect of feedback on user performance. It was found that feedback which contradicts users' expected model of ASR behaviour is detrimental to ASR use. In conclusion, dialogue for speech based interaction with machines should be short, succinct, and task specific. This will support operators 'natural' behaviour with ASR, and assist the recognition processes of the device.

Introduction

A problem often presented as a key issue for human factors of ASR is the design of adequate vocabularies for ASR. Although important in ensuring acceptable levels of performance, vocabulary design follows from the application domain. In other words, the vocabulary used to perform specific actions will be couched in the terminology of the domain. Thus, any guidelines will be general and must be reconsidered for each application and task.

We have seen, in chapter four, that ASR will work best if a vocabulary of phonetically distinct words is used, but this will not help if there are confusable words in the task vocabulary. One could redesign the vocabulary to make it amenable to ASR use, but this will force the operator to learn, not only a new means of interaction with the computer system, but also a new language. It would seem better to tailor existing language. Such tailoring requires an development of an adequate conception of how users ought to conduct their interaction with ASR systems. This will allow dialogues to be designed which are not only appropriate to specific domains but also to users' behaviour.

The question of what constitutes human dialogue with machines can be addressed from two fronts. In one, human conversation is used to provide principles and ideas for ASR (Leiser, 1989a). This, inevitably, overlooks the differences between human communication and human computer communication. The second approach studies how people actually interact with ASR systems (Richards and Underwood, 1984). This latter approach is used in the studies reported below. The data are analysed in terms of human communication, but the observations derive from 'speech based interaction with machines'.

When we use a computer, it is only when we have problems that we become aware of the separate roles of user and computer. Generally, we are only aware of performing a task with assistance of the computer. This assistance leads us to forget that we are actually communicating across very obvious system boundaries. It is clear that although computers display information on their screens in a form which humans can understand, the computer does not use this form of representation internally, i.e. it processes information in a different manner to the user. This processing requires the use of a specialized form of representation and means of manipulating this representation, embodied in the computer's program.

We use a computer to assist in the performance of a task. The 'dialogue' is a means to achieve the task, and is not the task in itself. When the task is being performed smoothly and efficiently, we may not be aware that we are using a computer so much as performing a task, i.e. we are aware of our actions in performing the task but not of the physical presence of the computer. 'Dialogue' is used as a metaphor for the exchange of information. When HCI is manual, this metaphor can only be very loosely applied. When HCI uses speech, we need to be very guarded in the assumptions we use. It is only when the task performance falters that we become aware of what Pinsky (1983) calls the "irradicably double nature" of HCI, that it is both "mechanical" and "conversational".

The flow of information between user and computer needs to be set in terms which the human can understand. This is the goal of 'user friendliness'. But these terms also support the illusion that user and computer are speaking the same language, when in fact communication occurs across a system boundary. Writers in HCI assume that the interface extends from inside the computer to inside the user. This again serves to foster the illusion that the interaction between user and computer is dialogic, that a shared symbol system is being used and that user and computer' understand' the same things. Shneiderman (1980) proposed that the goal of using unrestricted natural language should be replaced by dialogues using precise and concise language. The question of whether to include knowledge bases in ASR devices to assist with dialogue cannot be addressed until an acceptable model of human computer dialogue has been developed.

In control room systems, users should be encouraged to view the computer as an intelligent servant or as a 'cognitive tool', rather than as a dialogue partner. Users would then be using vocal commands to control the operation of the device. This appears to be very different style of vocal interaction to that humans are used to. However, there are situations in which humans communicate using limited vocabularies, and restricted syntax. Such situations are in issuing orders, or talking over the telephone. Both of these situations rely on task specific vocabularies, built around simple constructions. If we adopt these styles of speech as metaphor for speech based interaction with machines, then using a computer could be analogous to speaking to a servant rather than a conversational partner. Therefore, they will use short, simple words and concepts. This is evident in the studies on ASR use reported later.

Providing dialogue is kept simple, language problems requiring knowledge, such as the resolution of anaphora and deixis, will not be encountered. If the dialogue is made more complicated, users could attribute the device with too much intelligence. It is possible to achieve a level of dialogue at which the user will be able to use natural language and the device will not need additional knowledge. For instance, it is well substantiated that computer systems can be developed which fool the user into attributing human intelligence to the device on the basis of their interaction with it (see Weizenbaum's (1966) studies with his ELIZA program, and Boden's (1987) report of a 'paranoid' computer).

The Structure of Dialogues

We can define human dialogue as sequences of coherent utterances. The utterances themselves are structured and coherent in terms of semantics, and also in terms of the dialogue. "Formal" languages assume local structure and coherence, but do not necessarily require such factors at a global level. If ASR requires global structure and coherence for its dialogues, then it is essential to include complex knowledge bases to control the interaction. However, for the purpose of issuing commands, be they control room commands or data base enquiries, the ASR system can work with only local coherence. Thus, needing only limited intelligence.

We can look at the issues of structure and coherence on two levels: one with respect to the individual participants, the other with respect to the communication process. Ultimately, the former will require a detailed model of human language use which is beyond the scope of this thesis. Further, by emphasising human language capabilities we run the risk of misinterpreting computer behaviour. Therefore, this section will concentrate exclusively on the latter view: addressing structure and coherence of dialogues in terms of the communication process.

Longacre (1983) states that the distinctive feature of dialogue is that it involves a "sequence of speakers". This fact is central to the human ability to use speech to communicate. When we learn a language, we learn far more than the individual words of that language. We learn how to form them into meaningful units, and how to combine these units in order to communicate. Communication will inevitably require that both parties understand the language being used (although as the discussion of language above suggests, they need not "understand" the language in the same way), and that both parties accept the roles of speaker and listener. If both parties were to speak at once then neither could understand the other. The switching of roles from speaker to listener is referred as "the rhythm of dialogue" by Jaffe and Feldstein (1970), or more prosaically, "turn taking" by Sachs et al (1974). Such behaviour is crucial to the smooth development of the dialogue.

Researchers have often noted that all human languages, which involve vocal activity between two or more speakers, are characterised by periodic switches between speakers. In radio conversation, this switch is explicit in the use of the word "over" to terminate a speakers' turn. But in normal conversation is not so easy to predict when or why switches occur. It cannot simply be that there is a pause at the end of each speaker's turn, or that the speaker will always end with a question to signal the next speaker to talk. Such simple, unambiguous signals are not used solely to indicate turn termination, e.g. a speaker may pause for effect, or may use rhetorical questions.

Sacks et al (1974) develop a "systematics" for turn taking in conversation based on the notion of speaker selection. They note that, in conversation, speaker turns recur; that, although it is not uncommon for several people to speak at once, generally only one person at a time speaks, and multispeaker turns are very brief; the lengths of turns vary; transitions between different speaker turns are usually very brief, with little or no gap between turns, and occasionally with turns overlapping. By defining these rules, Sacks et al (1974) sought to define legal points of transfer between discrete preceding and succeeding points. To emphasize this, they use the notion of 'adjacency pairs'. These are characterised as dialogue acts which involve a turn by each participant. The most simple adjacency pair would consist of

A: "Hello" followed by B: "Hello"

But they can be extended to include questions and answers etc., even to the extent where several adjacency pairs can be embedded in one dialogue. The concept of 'adjacency pairs' follows naturally from the view of a dialogue as a sequence of events in time (Reichmann, 1985). Adjacency pairs show how the preceding discourse constrains, and can be used to predict, the development of dialogue. There are, of course, points at which they are violated, e.g. interruptions, or when they are suspended, e.g. jokes. But in most dialogues, they can be seen to operate.

In dialogues, a speakers' turn lasts as long as it takes to complete a conversational move. So, for instance, when someone begins an anecdote with what appear to be false starts or hesitations, these might, in fact, be attempts to gain the floor.

There does seem to be a set of cues which, taken together, contribute to effective turn taking. For instance, when a speaker self selects, she might take a breath or change posture or facial expression, or use some form of filler, e.g. "um","ah" to draw attention to the fact that she wishes to speak. Therefore, partners in a conversation must negotiate who is to speak and who is to listen.

These points combine to form a picture of dialogue as a process of creative problem solving, a means of constructing meaning through inference and action. So negotiation refers to all aspects of dialogue, from meaning to topic to turn taking. In normal dialogues, these combine into a very unwieldy body of evidence, which can make research very difficult. But there are certain types of dialogue which place restrictions on the participants. These can be seen to be more analogous to HCI than normal dialogues.

"Limited" Dialogues

Frankel (1984) noted that in doctor / patient consultations, the doctors speech is mainly carried out using questions. Indeed, he estimated that 90% of the doctors' utterances were questions. Thus, turn taking can be seen to be scripted, in terms of the dialogues setting. The patient will expect to be prompted to provide information concerning their complaint.

A further limitation one will expect from the use of ASR is that the channel of communication between participants will be highly limited. The use of extralinguistic cues, such as eye contact and body movements will be lost entirely. However, this is somewhat akin to dialogues carried out over the telephone, as can be seen from the observation of Rutter (1987) below,

> " Communication face to face is rich in social
> cues: we can see one another; ... Over the telephone,
> in contrast there is little except the voice..."

Rutter (1987) reports, a study carried out by Bell Telephones in 1970. Three thousand members of staff were asked " what would be a suitable alternative to face to face communication?" 85% insisted that they would need both sound and vision, although half of these decided that sound and some graphics facility would be acceptable. Only 2% of those interviewed accepted sound alone.

Fielding and Hartley (1987) point out that people speaking on the telephone tend to adopt a specific style of conversation, which can be characterised as 'short, sharp and sweet'. This means people are normally task oriented in their topic of speech, and stilted and nonspontaneous in their style. Jaffe and Feldstein (1970) suggest that it is because speakers on the telephone lack visual information that they limit their speaking turns to short phrases. Long stretches of silence are perceived as uncomfortable and disconcerting to telephone users, and must be filled by speech.

From this, one would expect people to speak to computerswith task oriented speech . One would further expect their speech to be "short, sharp and sweet" because of the limited amount of feedback available to them. They will not have, nor expect, the various 'extralinguistic' cues of normal conversation, such as 'body language'. This can be further illustrated by people's use of speech in problem solving exercises. An example of such research can be found in Chapanis (1975). Two subjects are required to solve problems using a variety of communication channels. One subject acts as information provider, the other as task performer.

Early studies, such as Chapanis (1975), often provided quantitative rather than qualitative data, e.g. in the Chapanis study, speech was shown to be as effective a means of communication as an 'communication rich' environment, but how the various instructions were structured or how the interaction proceeded was not considered. These studies only emphasise the aspect of speech to give commands. They tend to play down the important roles of speech to define the problem space, to name the components in such a way that both partners know what is being talked about, to generate a shared knowledge base, to cope with misunderstandings by redefining the knowledge space or renaming components.

Grosz (1977) shows that human dialogue in problem solving is affected by the nature of the task. There exist certain points in the dialogue at which a direct interaction between task and dialogue occurs. She calls these points foci. Shifts in focus were shown to correspond to shifts in attention between different objects. This notion of focusing is similar to the idea of the 'given/ new' concept which Halliday (1971) adapted from the Prague school of linguists. In this concept, 'given' information is that which both speaker and hearer can be assumed to share; 'new' information is that which the speaker needs to introduce into the discourse in order to develop it.

Krauss and Glucksberg (1974) investigated the social code of communication. By this they mean the development of a specialized jargon in the performance of a task. Their experiments used pairs of subject and a set of blocks identified by abstract designs. The task of the subjects was to place the blocks on a pole in the same order. One subject was to describe the blocks as she placed them on to the pole, and the other subject used these descriptions to place the blocks in the same order. Subjects were separated by a large board. Krauss and Glucksberg (1974) found that the initial descriptions of the abstract designs were very verbose and tried to mention as many aspects of the design as possible. But as the trials were repeated the descriptions appeared to become more succinct, until they were single words. It took about four trials for a description to go from nine words to one word.

The development of specialised forms of human communication can be found in such environments as Air Traffic Control. Controllers tend to speak to each other using specific languages, which Falzon (1984) terms "operative languages". These languages are constructed within the constraints of the task and its domain. They are characterised by short, simple phrases, often without any syntactic construction, and a highly limited vocabulary. Falzon (1984) argues that controllers merely need to use 'schema' words to communicate quite complex ideas. A 'schema' word is defined by the underlying schema or set of concepts which need to be understood in order to understand a spoken command. The development of such a 'schema' language can be observed in the Krauss and Glucksberg (1974) study. Subjects are able to develop suitable sets of concepts to describe an object, and these concepts can be communicated using one word.

Grosz (1980), in a study using expert / novice pairs, found that level of knowledge of a particular task influences focusing and how descriptions are formed, e.g. experts tended to focus on the functional aspects of the system as a result of their detailed knowledge of how the system worked; while novices tended to rely on perceptual features, such as colour or size of each component.

What these studies suggest is that there is no "right" way of carrying out a dialogue for the purpose of solving a problem. Rather such dialogues are the result of the exchange and development of a shared knowledge in terms which are acceptable to both parties. In the expert/novice experiment, there appeared to a negotiation as to what constituted a shared knowledge that both novice and expert could work with, this resulted in both parties introducing 'new' components using their perceptual features and then moving to a functional description of 'given' components.

Grosz and Sidner (1986) propose that in order to get someone to perform an action, the speaker does not necessarily require them to believe that their action will necessarily contribute to the desired outcome. It is sufficient that the speaker believes that they are capable of performing the required action.

Study Four

Initial users of Automatic Speech Recognition (ASR) devices are often prone to adopt a monotonic, robotic voice. It has been suggested that such behaviour is inappropriate to the use of ASR devices because when users controls an application using ASR, their style of speaking changes from that which they used during the process of template creation (enrolment).

Richards and Underwood (1984) studied users interactions with a telephone based information service. An experimenter acted as the information service and replied to subjects enquiries either through a normal telephone handset or through a vocoder. The nature of the task obviously affected the speech behaviour, making it less varied and more ritualised than ordinary conversations (as Barnard, 1974 also found). However, it was found that the type of feedback received affected the users' style of speaking. Thus, users speaking to the 'machine' spoke more formally and more slowly, using fewer pronouns or expressions which might require the listener to make any inferences, than those speaking to the 'human'. Users speaking to the 'human' were also observed to use more fillers in their speech, i.e. 'ahs' and 'ums', confirmations ,i.e. o.k., and 'polite expressions. These are used for the maintenance of the ongoing of the conversation, rather than for the purpose of giving information.

The differences between subjects talking to the 'human' and the 'machine' can be explained, in part, by Jaffe and Feldstein's (1970) notion of conversational equilibrium. They found that, especially in formal situations, such as interviews, speakers influence each others style of speaking, most noticeably in terms of pause length, number of pauses, length of utterance etc. One might expect such results to come from situations which involve quite high levels of uncertainty on the part of the conversational partners.

It could be argued that Richards and Underwood's (1984) study, allowed users the opportunity to use spontaneous speech, albeit in a limited task domain, because the experimenter was required to exhibit some intelligence when dealing with user requests. Currently available ASR systems do not exhibit such intelligence: indeed they are hard pressed to recognise the speech in the first place. Also, people are used to making judgments of a person's intelligence or social attractiveness etc. on the basis of the way that they speak (see Giles and Powisland, 1975). One might expect the harshness of the vocoder speech to effect the users responses to it.

In this study, the "Wizard of Oz" technique (Kelley, 1983) is used to study speech based control of a process control simulation. This technique is a valuable for human factors, in that it allows observation of user behaviour in systems which, due to current limitations, cannot be implemented (Small and Weldon, 1983). The proposed computer system is simulated by the experimenter, who mediates subjects commands via a function keyboard (see below) or a palantype keyboard (see Newell et al, 1987).

The task was kept simple in terms of complexity and possible vocabulary variation. Firstly, in an effort to reduce the the necessity of the experimenter to make any inferences. Secondly, because the study aims to look at speech style, it was felt that variation in words used should been kept as low as possible. Although subjects were given free choice in the vocabulary which they could use, the task required only a limited number of words.

10 undergraduate students (six men and four women; mean age: 20) acted as volunteer subjects in this study. They were assigned to either a 'computer' or 'human' group. The study took twenty minutes per subject. The task of the subjects was to use spoken commands to control a simulated process. The process control task was the same as the one described in Study One (chapter seven).

After receiving an explanation of the process control simulation, subjects were given a questionnaire to complete. This was not intended to collect data, but was used as a means of reinforcing the subjects' belief that they were taking part in an assessment of a speaker independent, ASR device. On completion of the questionnaire, subjects carried out the task of controlling the simulation. They were prompted with < say "Go" to start> displayed in the operation log. After saying "go", subjects were free to issue whatever commands they felt to be appropriate.

In the 'human' condition, subjects spoke commands to the experimenter, seated beside them. The experimenter then typed the commands into the function keyboard. Subjects were instructed to keep 'extra task' speech to a minimum. In the 'computer' condition, the experimenter was seated in another cubicle, listening to the subjects' commands and responding to them using the function keyboard. Thus, subjects were led to believe that they were interacting with a computer when in fact they were speaking to the experimenter. Subjects speech was recorded, in both conditions, and analysed in terms of number of commands, words, and unique words used. Statistical analysis employed a Mann- Whitney U test.

Results

	No. Commands	No. Words	Unique Words	No. Words /command
Computer	108.67	231.5	22.3	2.17
Human	83.5	254.33	33	3.06
Significance	p<0.05	n.s.	p<0.05	p<0.0001

Figure 10.1: Table of Results for Study Four

The "computer" group issued significantly more commands than the "human" group. The "human" group used more unique words and used more words per command than the "computer" group.

Discussion

The 'Computer' Group appeared to be less comfortable with silence in their dialogue, than the 'Human' Group. This is as expected, from the definition of 'dead time' in the use of radios, discussed above. When subjects cannot use extralinguistic information and contextual cues, they need to rely on speech. There were clear, significant differences between the two groups. Although both Groups used a similar number of words, the 'Human' Group used a significantly greater number of unique words, and more words per command. This suggests that their interaction was more conversational than that of the 'Computer' (see also Hauptman and Rudnicky, 1988).

The use of the 'Computer' led to an adoption of short, succinct speech. North and Lea (1982) also found that users prefer a succinct as opposed to lengthy type of command when using ASR. This style of speech is investigated in study five.

Study Five

Seven students (Six male postgraduates, and one female undergraduate) were paid £2 for participating in this study, which lasted forty minutes. The study replicated the 'Computer' condition of study four.

Although the questionnaire and assessment sheet used, were intended to support the "Wizard of Oz" deception, rather than as data collection instruments, it is interesting to note that the attitude of all the subjects to ASR changed as a result of taking part in this study. On the pretask questionnaire, they expected ASR to perform worse than a keyboard; on the post task assessment sheet, they rated ASR as being able to perform better than a keyboard. This suggests the favourable conclusion, that one only needs to introduce users to an ASR system to overcome any initial skepticism that they may have (see also the observations from study three, chapter nine). Obviously, one cannot read too much into these results, after all no effort was made to uncover the criteria subjects were using in their ratings. Also, the "ASR" system used in this experiment did not require the irksome task of enrolment, which might be expected to influence users' attitudes towards the system.

It would be interesting to examine performance figures in comparison with those obtained in study five. However, these three studies are intended to form a set, each examining different aspects of the speech based interaction with computres. Further, due to the small population used in this study, and the nature of the data, results will not be dealt with statistically. Rather they will be discussed within the discourse analysis framework outlined below, and will be presented as qualitative support for the observations which arise. We will regard dialogue as a progression of 'adjacency pairs' (Sacks and Schlegoff, 1973), bounded by initiation and termination expressions. Brown (1984) points out that these adjacency pairs are very flexible, and appear to progress using a standard path successor, i.e. the adjacency pair "question"/ "answer" will have a small set of plausible expressions for the answer part. The plausibility of the answers are defined by the relation to the question.

The fact that flexibility occurs, entails problems which must be dealt with to ensure smooth progression of the dialogue. When problems do occur then the dialogue is likely to undergo some sort of breakdown which requires dialogue participants to initiate a dialogue repair strategy to help them recover the 'gist' of the dialogue. By using this simple description of dialogue, we are able to make several observations of speech based interaction with the process control simulation. We suggest that these observations can be extended to cover all speech based, human computer interaction.

Subjects were prompted to begin the session with a prompt in the operation log box, which said < say 'GO' to start>. The operation log box was then cleared in response to a key press by the experimenter. Subjects were then free to begin issuing commands to control the process, generally the next commands were for a display of the output monitor's reading. One subject failed to see the change in the operation log box, and issued several other commands, such as 'begin process','set system running', before commanding "show output display".

Vocabulary Selection and Use

Once the session had started, the dialogue progressed according to the adjacency pairs formulation, as was expected. The pairs being 'command' (issued by the subject) and 'response' (from the 'computer'). The variation between commands, and the ways subjects used the responses will form the first part of this discussion. Then error handling strategies will be examined.

Despite the fact that this was a very constrained task domain, the subjects were faced with a number of options concerning the vocabulary which they could use. There were two main tasks, and I shall look at vocabulary selection and use under these tasks.

a.) there was the monitoring of the outputs of the units. This required the subject to give a command to call up the graph of the unit s/he wished to look at;

b.) there was the resetting of the output of a particular unit down to base level. This required the subject to issue a command to perform this operation.

a.) Output Monitoring

There was some variation in the commands used by subjects to call up the graph for the units. I had expected subjects to issue commands which took the form <verb><object>, as this seemed to be an everyday structure used to issue a command. (The object is obviously the desired unit and will be identified by its colour/number code). However, only three subjects used a <verb><object> structure. (Subject 7 used "Display [ ]" once out of 168 monitor commands (used to call up display of unit graphs); Subject 5: used "Check[ ]" 137 times for the 139 monitor commands he gave; subject 7: used "Display[ ]" 21 times out of the 163 monitor commands he gave). Only subject 5 can be said to use this construction consistently.

The other subjects used the construction colour/number. In using this construction, the use of a verb appears to be redundant, presumably because the subjects assume that the computer 'knows' that in identifying a unit, the subject is also requesting the graph of that unit to be displayed. Thus, they make the assumption that the computer and user share knowledge about the presence of the verb, and this shared knowledge makes the explicit use of the verb redundant.

It seems apparent that subjects will use the vocabulary which they is most appropriate to the task and to the capabilities of the machine with which they are interacting. This is supported by examining whether subjects used a word or a letter for the colour part of the unit code, i.e. did they say "red" or "r". In most cases (i.e. except for subjects 1 and 7), the nature of the colour identifier (word or letter) was determined by the subject's first utterance: if they said a word first, then they would persist in using a word throughout the task, and similarly with letters. Subjects 1 and 7 showed changes in their selection of colour identifier: subject 1 started using a letter and after 5 minutes switched to using a word; subject 7 tried both but letters and words used words more consistently.

Subjects were questioned, after they had completed the assessment sheet as to whether they had used letters or words. Subjects (2, 3, 4, and 6) who used words said that they did so because the letters "might sound the same". This was also the reason given by subject 1 as to why he changed from letters to words. In other words, the command sequence a subject uses is the least complex and shortest one, which is least likely to lead to confusion, and which achieves the desired goal.

Subjects who used letters said that the computer was using letters to describe the units, i.e. in the operations log. Subjects constructed their vocabulary around what they considered to be shared information, and in what they considered to be shared representations of that information.

There was a relationship between whether the subjects used words or letters and which part of the screen they used as their main source of feedback. Subjects were asked to rank the four boxes on the screen in terms of amount of useful information, and in terms of amount of time used. Not surprisingly, boxes which were seen as conveying the most information were used the most. But it was also found that subjects who used the box containing the graph tended to use words, and those who used the operation log tended to use letters.

Subject 1 said that he had used the operation log box to check that what he had said had been correctly recognized, until he realised that the graph also showed the unit's identifier. Then he just looked at this box, to save having to look at two places on the screen for information. Subjects construct their vocabulary according to their perceptions of the system's ability to correctly interpret what they say, and according to the ease of use. These observations are supported by similar studies carried out by Zoltan Ford (1984). She used the ' Wizard of Oz' technique to study peoples' behaviour using a speech operated database enquiry service. She allowed the users unrestricted speech input, i.e. they were free to choose whatever commands they thought necessary and to phrase the commands in whatever style they thought appropriate. The 'machine' replied using a restricted vocabulary and simple syntax. It was found that users quickly adapted to the style of interaction which the 'machine' was using. This is further investigated in study six, below.

b.) Resetting Plant Deviations

When the experimenter described the task to the subjects, care was taken not to use just one word to describe the activity of resetting the output of a unit. This was intended to reduce the possibility that subjects would adopt this word rather that selecting their own.

All subjects, apart from subject 2, used a word other than reset (the word which the operation log displays in response to the resetting command) to perform the action of resetting the output of a unit. Whether they persisted in using this command or whether they adopted "reset" depended again on which box they used for feedback. However, this is not necessarily related to the box they used for feedback to check recognition, or which determined whether they used a word or a letter. Although a correlation between feedback and vocabulary has been suggested, subjects who use "reset" are not necessarily those who used a letter for the colour identifier. It is suggested that subjects require different types of feedback to provide information concerning different types of system performance (these points are developed in chapter eleven).

During debriefing, subjects were asked how they knew that the resetting command they had given had been recognized. Subjects 1, 2, 3, 4 and 5 said that the computer "said reset in [the operation log box]"; subjects 6 and 7 said that the graph went down. This suggests that subjects preferred to check that the resetting command had been correctly recognized as a word before accepting the performance of the resetting action as evidence of recognition.

It was also apparent that subjects had difficulty thinking of an appropriate word to use for the resetting command, there being no information on the screen to help them. Subjects 1- 5 all paused for a significant period of time (i.e between 10 and 15 seconds) when they were first in a position to give a resetting command, and they offered a word quite tentatively, e.g. reduce, down,set to zero etc. After this they tended to use reset. Subjects 6 and 7 persisted in using variants of the command they had first used for around 5 minutes (i.e. one third of the total task time) before they adopted 'reset'. During debriefing they both said that the movement on the screen made them look there rather than in the operation log box, and that this movement was taken as confirmation of recognition.

Error Correction

Errors were deliberately introduced by the experimenter at random points in the experiment. These errors took the form of either substituting a colour for the colour spoken, or substituting a letter for the letter spoken, or substituting "Red Six" for "Reset". When faced with such errors, subjects tended to repeat the command with a pause between the words and stressing the misrecognized word. If the misrecognition persisted to this attempt at correction, then subjects tended to repeat the command with a pause between words and stress on both words. Subject 5 corrected errors by repeating the command without the verb, and stressing the misrecognized word. Errors are held to arise from the systems imperfect 'hearing' of the subjects' speech, i.e. errors are of an acoustic rather than linguistic nature.

Study Six

It was suggested, from study five, that the speech of ASR users will be effected by the type of feedback that they received. Ringle and Bruce (1982) pointed out that human dialogue progresses according to certain limits. Dialogue partners generally adapt the style and content of their speech in line with such features as topic of interest, complexity of vocabulary and syntax, and level of attention. This produces a dialogue framework which accommodates both dialogue partners and makes communication easier.

We have already mentioned the work of Zoltan Ford (1984), in which subjects were observed to modify their speech style to fit in with the restrictions imposed by a simulated ASR device. Ringle and Halstead Nussloch (1989) also demonstrate that subjects adapted their response to a 'formal' form of feedback, in line with system constraints. Leiser (1989b), in a similar study, showed that subjects' typed queries to a database began to employ terms and structures used in the 'computer' feedback. This shows that convergence can occur in human computer dialogue, and further supports the observations of Krauss and Glucksberg (1974), in which subjects developed a form of 'jargon' in the performance of block assembly tasks. Study seven was designed to consider the question of convergence in speech based interaction with machines.

Ten undergraduate students (five men and five women) acted as volunteer subjects in this study. The study lasted for twenty minutes. Subjects were divided into two groups: verbose and limited feedback. The method used was identical to that of study five, except that the two groups of subjects received different types of feedback in the "operation log". For the 'Verbose Feedback' group, the identification of a particular unit took the following form:

<Red Nine Called>

For the 'Limited Feedback' group, the identification of a particular unit took the following form:

<R 9>

Results were analysed, in terms of number of unique words and number of words per command, using an independent measures t test.

Results

	Commands	Words	Unique words	No. words x command
Limited	110. 54	239.87	22.16	2.17
Verbose	107.6	268.7	16.50	2.50
significance	n.s.	n.s.	p<0.05	n.s.

Figure 10.2: Table of results for Study Six

It is interesting to note that although the limited feedback group issued more commands than the verbose feedback group, the verbose feedback group used more words and had more words per command. Unfortunately, these results are not significant. However, we can say that the "Limited" feedback group used significantly more unique words than the "Verbose" feedback group (p< 0.05).

Discussion

Study six failed to replicate the findings of Zoltan Ford (1984). It did, however, support the hypothesis that feedback will effect user performance. There is an indication that the more feedback subjects recieve, the more words they will use. However, this cannot be substantiated as the data are significantly different. One can say that the data in figure 10.1 are similar to those in 10.1, which suggests some consistency in the way people speak to computers.

There is a significant difference in the number of unique words subjects used, and this could be assumed to relate to the subjects perception of the computer they were speaking to. The two types of feedback were different in the information that they provided the subject. While the 'Limited feedback' provided simple feedback related to a particular object, the 'Verbose feedback' also related to the action required. From the preceding discussion, we can see that the 'Limited feedback' capitalised upon the assumed notion of an intelligent servant. The feedback simply echoed the subjects command. The 'Verbose feedback', however, could be likened to a conversational partner; the information relating to the choice of action suggests the possibility of negotiation. This, in turn disrupts the conception of ASR interaction which we suggest users develop, and hence, leads to an inhibition of user behaviour. The degree of uncertainty presented by 'Verbose feedback' hampers subject performance, which suggests that users of ASR require feedback which provides a clear indication of what the device has recognised, and what the system is doing (see chapter eleven).

One cannot, and need not, argue that this disproves Zoltan Ford's (1984) findings. Rather, her study used a word processing type task. The only feedback that the subject received was a line of text. Consequently, variations in this feedback affected the subject's interpretation of the ongoing dialogue. In study six, on the other hand, subjects receive feedback from several sources (see Study One, chapter seven). In addition to the "Operation Log", there is the "Unit/Monitor Display", the "Alarm Panel" and the "Plant Diagram".

This would suggest that the extra information in the "Operation Log" will present a hindrance, rather than a help, to the user. It is suggested that the textual feedback the user received related to their conception of ASR performance. In other words, textual feedback, in the "Operation Log", told them how well the device recognised what they had said. Thus, all that is required is an 'echo' of subjects speech, in terms of the task. This 'echo' need not be verbatim, but simply state the name of the object selected. This further supports the idea the dialogue in ASR should be short, succinct and task specific.

It also raises the interesting issue of forms of feedback for ASR in control room systems. While feedback concerning recogniser performance may be best presented using limited text, the user still needs to be told about plant activity and performance. The question of how best to provide feedback of these different measures is addressed in chapter eleven.

Model of Dialogue in Speech based HCI

Consider an analogy between using ASR and talking to someone over a noisy telephone line: if they cannot hear what you have said correctly, you do not, usually, assume that this is a result of their imperfect knowledge of the language, but because of interference on the line. Subjects had all their commands promptly responded to and, to all intents and purposes, the "ASR" system was as competent a user of language as they were, within this limited task domain. Therefore, any errors would result from it not 'hearing' correctly rather than its lack of linguistic knowledge.

It might be suggested that one could offer the analogy of talking to someone who did not understand the language one was using. However, it is typical in this circumstance to speak slowly with equal stress on each word so as to over-emphasize what one is saying. If this was the model subjects were using then one would expect their speech to be slower than normal and with even stress.

The method of error correction described above suggests that the subject's model of the system has more in common with subjects' experience with other speech perceivers, i.e. humans, than with computers. This is supported by an analysis of the use of stress in the "monitor commands" used to identify units.

When the command to display a new units' output was given, the position of the stress in the utterance depended on whether the colour or the number had changed from the previous display. If the colour had changed (irrespective of a change of number), i.e.. " Red Seven" to "Blue Twelve", then the stress would be placed on the colour. If it was the number which changed, and the colour stayed the same, then the stress would be placed on the number. It was quite easy to pick out which of the two words the subjects was stressing. This relation holds for around 80% of these commands.

Given a choice of vocabulary to perform this task, subjects tried to construct short phrases. In some instances, subjects could be seen to actually shorten the phrase they used, i.e. subject 7 began with " Display[ ]", then used colour(word)/number, then used colour(letter)/number, before resorting back to colour(word)/number.

Study four showed that people speak differently to a person sitting beside them than to a 'computer'. But suppose they knew that the 'computer' was in fact a person? If, in this instance, their speech was still highly limited (as in the 'computer' group), then our results could be described as an artifact of the constrained task domain this experiment used.

In a pilot study, for study five, subjects were used who knew that the experimenter was acting as a computer. Even though subjects attempted to act 'naively', they seemed to be aware of the experimenter's presence, as the use of extraneous speech, e.g. "o.k.", "thanks" etc., showed. In this study, subjects did not use speech which was unrelated to the performance of the task in hand nor did they use 'fillers', i.e. 'ums' and 'ahs' which were prevalent in the pilot study. That is they did not need to keep the interaction going with dialogue control expressions, as one finds in human conversation.

One might expect subjects to keep commands as short and as relevant as possible for two reasons. Firstly, so as to minimize the demands on their attention and time. Secondly, in line with Grice's (1975) maxims of brevity and relevance.

This, together with the result of Studies five, six and seven, suggest that rather than treating the 'computer' as a "dialogue partner", as some writers in HCI claim, the subjects treated the 'computer' as an "intelligent servant". This notion was suggested by one of the subjects who went on to say, "you tell it what to do and it gets on with it. You don't need to say anything else or have some sort of chat or something". Given that the 'computer' is an "intelligent servant", one wonders whether the sequence of commands are perceived by the subject as existing as part of a continuum over time, i.e. as a conversation, or as isolated events, ie. as subtasks of the overall task of controlling the process control simulation. The evidence from the use of stress in the "monitor commands" could be taken to support either supposition. If the interaction was the result of discrete events, isolated in time, then the marked information would represent "new" information. If the interaction was dynamic then the marked information would represent "inferable" information.

Even though the process control task was designed to be a dynamic task (and was described as dynamic to the subjects), this does not necessarily mean that they will see each command as part of an overall task of controlling the plant, rather than as issuing a command as a task in itself. One would suspect that because the plant diagram is given on the screen, the subjects would assume that the 'computer' possessed knowledge concerning the identity of the units. Therefore, the marked information would convey "inferable", as opposed to "new", information. On the other hand, the fact that 'Verbose feedback' actually hampered performance would suggest that subjects did not regard the interaction as a dialogue with a 'dialogue partner', but as a process control task which required spoken, as opposed to typed, commands. The 'dynamic' aspect of the task would not relate to an ongoing spoken dialogue, so much as an ongoing task requiring discrete commands to be issued. This, again supports the notion of the short, succinct, task specific style of dialogue, as opposed to the conversational, in ASR use. As well as supporting the 'natural' behaviour of users, this dialogue style can be recommended because it will not over burden the 'cognitive resources' of the users. That is, because the users are not involved in a conversation, they do not need to maintain a record of previous commands, in memory, as a conversation requires. Rather, each command is spoken in isolation, and in relation to the process or task in hand.

CHAPTER ELEVEN

FEEDBACK FOR ASR IN CONTROL ROOM SYSTEMS

Feedback is a very difficult term to define; it means many things to many people. However, from a review of related literature, it is possible to offer a set of working definitions. These definitions are then applied to the many instances of feedback proposed for ASR use, with especial reference to the control room. It is argued that auditory feedback, in the form of speech synthesis, will be unacceptable for a number of reasons. Therefore, feedback should be presented visually. The definitions of feedback for ASR use lead to the conclusion that operators require two types of feedback: one relating to ASR performance, and the other relating to the performance of the system they are controlling. Study seven shows the feedback relating to ASR performance should be presented using text. However, study eight finds that the use of textual feedback in a process control task is prone to user errors. Therefore, it is concluded that feedback should not only be relevant to the task in hand, but also be integrated in the main display .

Definitions of Feedback

Most ergonomists would probably agree with Norman (1988) in defining feedback as the practice of,

> "sending back to the user information about what action has actually been done, what result has been accepted".

The user performs an operation and, by this defintion, feedback provides knowledge of the results of that operation. The operator must interpret this information, and act upon this interpretation. The information contained in the feedback message can relate to current system state, or to previous states, requiring assimilation of the information into a model of plant process (see chapter one). Feedback can seen to be a means of controlling behaviour. It can limit the effect of errors by informing the user of ongoing activity. It can be used to reinforce appropriate behaviour in the user. This illustrates that the term 'feedback' has several meanings, and can have several uses.

Like many terms used in human factors, 'feedback' has been borrowed from a discipline where it has a precise meaning. Psychologists tend to use feedback to refer to any information concering the results of a person's behaviour which subsequently affect that behaviour. The ways in which this information is used is often not studied; 'feeback' is thus an assumed process.

Moray (1981) has criticised the "extended definition" of feedback in psychology, and reminds the reader of the definition of a closed loop negative feedback (CLNF) system in control theory:

> "A system is said to contain a feedback loop when the output of that system interacts with its input in such a way as to modify the subsequent activity of the system as it continues to generate output".

In a CLNF system, a desired output level is set. Maintaining this output level can be thought of as the 'goal' of the system. Any deviations of output from this level are fed to a controller in the form of an error signal. The controller then modifies the system function in order to maintain the desired level of output. This system allows prediction of system performance. Disturbances will cause proportional error signals, which in turn affect the controller, and can be dealt with by simply subtracting the output level from the desired level. A CLNF system can be seen as a dynamic, self regulating system operating continuously on a variable input.

In cybernetic theory (Weiner ,1948), human behaviour can be modelled by servomechanisms, similar to CLNF systems. Such modelling was thought to provide mathematically precise descriptions of human behaviour, for instance, operators of process control systems were considered to behave in a similar fashion to servomechanisms (Hickey and Blair, 1958). Thus, an initial definition of feedback, for human behaviour, can be proposed which describes the effectiveness of an action in reaching a desired goal.

In terms of simple tracking behaviour, the analogy between human activity and CLNF systems appears to be adequate. If subjects are required to move a pointer between two points, it could be argued that they continually sample the position of the pointer in relation to the goal state (Fitts, 1964). However, such an explanation confuses two basic types of feedback. Smith and Smith (1987) state that homeostasis and CLNF systems are both examples of intrinsic feedback. The feedback signal is generated within the organism or system. On the other hand, tracking tasks, such as those used by Fitts (1964), are examples of extrinsic feedback. The feedback signal is received from the environment.

While intrinsic feedback allows a simple decision to be made in terms of output levels, extrinsic feedback requires a more complex form of decision making to be made in terms of desired goal states. It is more usual that tasks will make use of extrinsic feedback, because it is generated by the equipment that the operator is using. Further, people are not always aware of the effects of intrinsic feedback upon them. This difference in level of awareness also allows a distinction to be made in terms of sampling rate between the two types of feedback. While a CLNF system will continuously sample intrinsic feedback, such as the self generated error signal, a human operator will only sample extrinsic feedback, such as an error signal from a machine, intermittently.

Uses of Feedback

Although he was initially attracted to the idea of human operators as servomechanisms, Craik (1947) realised that there was an essential difference. He had suggested that tracking behaviour could be modelled in terms of the time nerve impulses take to travel along a chain of synapses. This would assume a smooth, constant rate of performance. When people perform motor tracking tasks, they tend to sample their performance intermittently. This leads to large fluctuations in their performance, as they overcompensate for deviations. These results could be taken to suggest that feedback could be more usefully provided at set times than continually.

Craik (1947) noted that people tend to try to predict system states (see also Harvey, 1988). Therefore, the aim of controlling any system is not necessarily to compensate for deviations in the present state, but is directed towards the control, or attainment, of a future state (Kelley, 1968). Extrapolation and prediction have no place in closed loop systems, as both activities require the generation and assimilation of evidence beyond the current state.

Therefore, feedback is more than the knowledge of results of an action. It requires the interpretation of such results in terms of current system state. Balzer et al (1989) distinguish between 'outcome' feedback and 'cognitive' feedback. 'Outcome' feedback provides users with knowledge of the results of their action. 'Cognitive' feedback refers to information concerning the users' knowledge of the domain, the system, and their actions. 'Outcome' feedback would describe the use of temporal relations to model system performance; 'cognitive' feedback would describe the construction of evidence and heuristics to predict system performance.

Murrell (1976) points out that early system designers assumed that displayed information should allow users to read off exact values concerning system states, that is, provide 'outcome' feedback. However, there is some controversy concerning the effectiveness of outcome feedback on performance (Brehmer, 1980). Hoffman et al (1981) showed that cognitive feedback gave greater improvement in performance than outcome feedback, in an experiment in which subjects controlling dials were given different types of feedback.

Murrell (1976) lists nine uses of dial displays. These include not only reading off the value, but also checking for deviations, and comparison with other display values. They can provide information for warning signals, indication of status, and tracking of trends. In each use, the information read off the dial will be utilised in a different fashion. Therefore, it is important that one consider what the information contained in a display is to be used for.

Types of Feedback

It is an axiom of human computer interaction (HCI) research that in using a computer system, users should be informed of the effect of their actions (Smith and Mosier, 1984; Shneiderman, 1987). Holding (1965) distinguishes around sixteen different types of feedback, varying along dimensions of time, origin, and style of presentation. It has been stated above, that feedback can either originate in the system or organsim (intrinsic) feedback, or it can be received from the environment (extrinsic). Smith and Smith (1987) divide extrinsic feedback into a further three types which are termed reactive, instrumental and operational:

> "In the familiar operation of handwriting, the writer gets reactive feedback from his own movements, instrumental feedback from the movements of the pen...and operational feedback from the written words".

Reactive feedback refers to the feedback an operator can receive from controls. When using a keyboard, reactive feedback is provided kinaesthetically: the user can feel the key move. Thus, feedback can be seen to be inherent in the use of the device. Seibel (1972) notes that process of making a movement and striking a key provides a major source of feedback for experienced typists. This type of feedback is not normally noticed in using a keyboard, unless some malfunction occurs, such as a key sticking, which interrupts the smooth flow of use. This type of feedback can be considered as the lowest level of feedback. It is immediate, in that it occurs exactly when the device is used; it is inherent in the device, in that it results from movement in response to the user's action; and it requires minimal cognitive processing by the user, in that he often does not notice such feedback until a malfunction occurs.

When one uses a touchscreen, the actions performed are superficially similar to those of using a keyboard: one selects an area and touches it. However, while the movement of a key provides tactile feedback to the user, there are no moving parts on a touch screen. Consequently, it is customary to provide some artificial form of feedback to the user; either in the form of a tone or a highlight on the screen or both.

This type of feedback can be considered at the same level as reactive feedback in that it is immediate. However, rather than being inherent in the device it is added by the system designer. This means that some cognitive processing is required in that this form of feedback employs a simple code. The user needs to monitor the feedback signal in order to ensure that the area selected is the one chosen. Thus, artificial, immediate feedback can be assumed to intervene in the performance of the primary task by placing some cognitive processing requirments upon the user.

The second level of extrinsic feedback, defined by Smith and Smith (1987), is instrumental feedback. This is what Schurick et al (1985) term primary or task feedback. If the user presses a key on a function keyboard labelled 'Reset ', then she expects the graph displaying the output from a particular unit to be reset to a predefined level ; if I turn on a cold water tap then I expect cold water to flow from it. This type of feedback pairs an action by the user with an action by the system.

The third level of feedback concerns the information supplied to the user relating to the results produced by an an action. In the simplest example, pressing a key labelled 'a' will cause an "a" to be printed on the screen. If the user issues a command to 'open a file', then operational feedback will be provided to signal to the user that a file will be opened. The user could type 'open x' and the command would appear on the screen. Hitting return would cause the file to be opened. This level is termed operational feedback because it requires some decision to be made by the user concerning its validity, i.e. does he want to open file 'x' or another file; is the computer going to open 'x' etc.

Rosinski et al. (1980) compared the performance of typists, of different ability, with the presence of visual feedback. They found that the provision of visual feedback had no effect on the input speed or number of errors across the groups. However, the greater amount of visual feedback that was provided , the easier subjects found the task of correcting errors. This suggests that where errors are possible, users should be provided with enough feedback to allow them to detect and correct the errors. Therefore, feedback can be used both for data entry verification and for error correction.

In a study reported by Fitts and Posner (1967) three groups of subjects a sixteen item, choice reaction time test. The groups varied in terms of the feedback they received. One group received no feedback, a second group received feedback concerning the speed of their response, and the third received feedback concerning the accuracy of their response. The results showed that the group receiving speed feedback were faster but less accurate than the group receiving accuracy feedback, and that the no feedback group performed at a lower level than either of the other groups. This suggests that the type of feedback provided will effect specific types of performance.

Holding (1965) defined a dimension of feedback which showed the temporal relation of feedback to the task. It can be immediate or delayed. A second temporal dimension which can be used to describe feedback, defines it as being either concurrent or terminal (Smith and Smith, 1962). Concurrent feedback was shown to be the most effective in terms of performance and learning speed. Smith and Smith (1962) also show that separate item feedback should be more effective than accumulating the results of several trials. Finally, feedback can be continuous or intermittent. Smith and Smith (1962) show that continuous feedback is more useful than intermittent feedback. These points can be taken to recommend that feedback should occur as near to the performance of the operator's action as possible. This relation will be determined by the type of task being carried out.

Feedback in ASR systems

Craft (1982) described a parcel sorting device controlled by speech. He argued that because the vocabulary only required the digits 0-9, feedback was not necessary in this sytem. Unfortunately, he did not describe the use of the system in enough detail for readers to gauge its success. However, it is difficult, given the definitions of feedback generated above, to imagine any system performing without some form of feedback to the user.

Consider parcel sorting applications as described in chapter two. The operator selects a parcel, reads the code on it, and the parcel moves to the correct line. This is an example of action feedback; but in order for the operator to check that the parcel is moving to the correct line, he needs to decide which line it should go to. This is no different to manual parcel sorting. Alternatively, the operator could read the code into the device, check that the code has been correctly recognised using operational feedback, and let the computer make the decision as to which line the parcel should be sent. The operators task is then to enter and confirm the data, which offers a potential to speed up parcel sorting. Thus, whichever form the parcel sorting task takes, the operator will require a suitable form of feedback.

A study by Williges et al (1986) found that the use of feedback increased recognition accuracy from 65% to 97%. Schurick et al (1985) demonstrate that using feedback can improve performance using ASR by around 27% (from 70% recognition accuracy to 97%). A study by Poock et al (1983) found that if users were trained to use an ASR device without feedback, and then were given feedback during use, their recognition accuracy increased by 5%. Similarly, if subjects given feedback during training were not given feedback during use, their recognition accuracy decreased by 5%. Therefore, feedback appears to be important in the use of ASR (Martin, 1976; McCauley, 1984; Hapeshi and Jones, 1989).

Chapter one describes why ASR devices will never consistently reach 100% accuracy. Feedback will allow users to check their input commands, and make decisions concerining the devices responses, i.e. do errors occur which need correction. This means that feedback and error correction are inseparable.

Reactive feedback in ASR systems

In terms of reactive feedback, there is no inherent feedback in speaking into a microphone (Berman, 1986). Thus, some form of artificial feedback is required at this level. This could take the form of providing an auditory or visual 'beep' after each word has been spoken. Mulla (1984) used an auditory 'beep' to pace subjects speech. The device used in his study provided a memory buffer between the microphone and the ASR device. The timing of the 'pacing beeps' meant that subjects could actually speak faster than the device could recognise. However, it is not certain that it is either useful or desirable to encourage subjects to speak at a rate faster than a devices capabilities, what would happen in a stressful situation, when subjects speech rate increased further? They would assume that, as they could already speak faster than the device could recognise, their faster speech would not be a problem.

Whilst Karhan (1987) shows that feedback using auditory 'beeps' helps inexperienced users of ASR pause adequately between phrases when using a telephone enquiry service, McCauley and Semple (1980) found that in periods of high workload, users tend to speak more quickly and omit pauses in their speech than in normal working conditions. Thus, the use of auditory beeps could lead to more problems than they actually solve. As Martin and Welch (1980) point out, the user should not be expected to pay attention to a signal announcing that the device is ready for every word, rather the user should be able to provide a suitable period of silence between words in order for the device to function correctly.

The speech based process control task described in study one, chapter seven employed a 'beep' to indicate that a word had been accpeted. Subjects were required to wait until the beep occured before they spoke the next word. Whilst one or two subjects were able, with a little practice, to time their speech to match the beeps, the majority found the beeps distracting, irritating. and confusing. Subjects had difficulty in deciding whether the beep indicated that a word had been recognised or whether it was a prompt to speak the next word, thus auditory beeps also slowed down the interaction.

It was apparent that the beeps were providing the wrong sort of information to the user. They indicated that a word had been recognised, but not which word nor whether the device had mistakenly recognised some spurious noise for a word. A function of feedback is to reduce uncertainty on the part of the user. One can see that using an auditory beep as feedback can, in fact, heighten uncertainty.

A final problem with the use of auditory beeps is that they can only be used on isolated word ASR based tasks. If the device in use provides connected ASR then it is difficult to propose where a beep should occur; after each word or after each phrase or after periods of silence ? The same problems will exist for 'visual beeps', i.e. a flashing light appearing on the screen each time a word has been received.

Whilst some form of immediate feedback is required in ASR use, auditory beeps do not appear desirable. The use of such beeps could be considered to indicate a form of misrecognition. 'Beeps' are currently used on keyboards to indicate that the operation a user is trying to perform cannot be carried out by the computer. If an ASR device can not find a suitable match between the input word and its stored templates, then this nonrecognition could be signalled by a beep. However, in control room operation, the operator will need to check the feedback before confirming a command. This will ensure system safety and security. It will also mean that 'beeps' to indicate that the device has not recognised a word will be redundant. This is not to say that soem forms of redundant feedback may be desirable, but that 'beeps' are not.

Instrumental Feedback in ASR Systems

Where ASR is used to issue a command which has a specific action associated with it, then instrumental feedback could be beneficial to users. Such feedback would be immediate (providing system response times are acceptable), and would not require much processing by the user. In the telecommand demonstration, described in chapter six, it is possible for users to call up pages of information, e.g. if the user issued the command "overview", a display of the grid map would be displayed. The change of dispaly provides immediate, instrumental feedback.

It does not seem necessary to check which word has been recognised if the appropriate action has been performed; users make the pairing between word and action to assume that for action x to be performed, word x must have been recognised. However, some users may require both word and action in their feedback.

In the process control demonstration, described in chapter four, one of the tasks of the subjects was to reset the output of various units. A graph indicated unit output against time. Issuing the command reset brought the graph back to zero. Thus, instrumental feedback could be received from the movement of the graph. It was noteworthy that not all the subjects used this feedback. Some preferred to check that a word had been recognised first and then check that the graph had moved (see study five, chapter ten).

There is degree of uncertainty between issuing a command and receiving instrumental feedback. Subjects switching between pages on the telecommand demonstration (chapter six), were often surprised when the wrong page appeared (such surprise was not evident when the wrong word appeared in a recognised command). There appears to be the assumption that if the system performs an action, it has correctly recognised the command. The surprise results from an action being performed inappropriate to the command. Thus, although redundant in information terms, operational feedback would supplement instrumental feedback to reduce this level of uncertainty.

If the only information a user has concerning how well a command has been recognised is the performance of an action, it is difficult to suggest an efficient means of correcting any errors that occur. In the instrumental feedback situation, the user can only repeat commands until the desired action is performed. Naturally, in some situations this behaviour could be costly and so some form of confirmation should be elicited from the user before the command is acted upon. This confirmation will require operational feedback.

Operational Feedback in ASR Systems

Operational feedback is often considered as the sole means of providing feedback in speech based interaction with machines. The interaction is centred around the use of spoken language, and is considered in terms of verbal feedback. There are a number of types of feedback at the operational level . These can take the form of synthesized words or phrases, or textual feedback using words or phrases.

i.) Auditory Feedback

One of the major claims for using ASR is that it liberates the user from the VDU (Lea,1980). This allows the user to perform tasks in other areas of the workplace. It is argued that providing visual feedback will confine the user to the VDU and hence, remove one of ASR's potential benefits (Simpson et al 1985). If one is to use ASR as a 'hands free/ eyes free' medium, then the provision of auditory feedback is desirable. There is also the proposed advantage of auditory feedback, over visual, derived from the theory of Wickens and his colleagues, discussed in chapter eleven. This suggests that more effective data processing by the user will occur if stimulus and response are paired in the same sense modality, i.e. manual data entry with visual feedback, or verbal data entry with auditory feedback. However, there are a number of problems with using auditory feedback.

Auditory feedback has already been discussed in terms of reactive feedback above. It has been suggested that feedback should follow input with as short a delay as possible. In the case of using isolated word ASR, auditory feedback could be used to echo the users speech. This can be distracting and intrusive, with users being confused by the echo as to what they need to say next (Martin and Welch, 1980).

Welch (1977) and Berman (1984) point out the essentially transitory nature of speech. This means that speech must be attended to immediately or the information it contains will be lost. This places a burden on the user in terms of information porcessing, which is not found is visual displays. Further, there is the possibility that using such feedback could lead the users to try to mimic the synthesised speech. This would cause their speech to drift markedly from the recorded templates. There is evidence from social psychology that when speaking to someone with a strong regional accent, people often adopt some of the speech patterns of this accent (Giles and Powisland, 1975). This is especially true in situations involving some degree of stress or uncertainty, such as interviews. For novice users of ASR, there is a tendency to assist the device by speaking to help it (Nye, 1982). The use of synthetic speech will only exacerbate this (Leiser et al 1987).

Speech is organised temporally. This means that the information contained in speech is sequential. By providing word for word feedback, the system is placing an extra demand upon the user. Not only must he monitor the feedback to check for, and remember, misrecognitions, he must also construct the whole command to check that it is correct. Hapeshi et al. (1988) report a study on the effect of modality of feedback (auditory vs. visual) on subjects' ability to recall data entered. Concurrent feedback, in either modality, increased recall errors. This led to the conclusion that feedback should be provided after whole phrases have been entered. Little difference was found between feedback modality and recall when feedback was given at the end of phrases. Jones et al (1990) demonstrate that auditory feedback after each item is significantly more disruptive to performance than feedback provided visually by text or symbol.

Witten (1982) recommended that auditory feedback be provided after whole phrases in order to retain the sequential aspect of speech, and provide for a more 'natural' interaction. However, Ito et al (1989) show that if subjects are given whole phrase feedback, they tend to wait until the whole phrase has been completed, even when they understand what the phrase means and are thus in a position to interrupt. This means that such feedback could actually hinder performance in terms of speed.

Studies by Pisoni and his colleagues (Pisoni,1982), suggest that subjects are able to process the information relating to surface structure of sentences, e.g. whether or not the sentence contained a particular word, using synthetic speech rather than normasl speech. But natural speech gives better performance on high level questions. As Pisoni (1982) points out,

> "This pattern suggests that listening to synthetic speech somehow distracts the subjects' attention to the sound structure of the speech signal rather than to more abstract levels of linguistic analysis that are involved in computing the meanings of sentences."

Further evidence of the effect of sound structure on the processing of synthetic speech was found in a short study carried out with a class of 32 undergraduate students at Aston University. Subjects heard the ICAO alphabet produced by a synthesis by rule device. They had to record what they thought it said. Results showed that they were 40% accurate in their recordings. Similar sounding words were often substituted for ICAO words, and where no word was recorded, subjects said that it did not sound like any words they knew.

When given a sheet containing the words of the ICAO alphabet, subjects could recognise synthesised speech with an accuracy of 95%. In this latter condition, subjects were required to match the synthesised word with its expected sound. These results suggest that auditory feedback would be useful as a means of checking individual words spoken, but not if judgments were to be made at a deeper level, e.g. users could confirm each word in a command but not necessarily the whole command.

Waterworth (1984) found that subjects could use auditory menus for simple, highly restricted information, but when it came to more complex information, subjects found the structure and quantity of information difficult to cope with. This would force subjects to concentrate on the surface structure. Luce et al (1983) found that natural speech gave better recall than synthetic speech. This difference was emphasised as task difficulty increased.

Schwab et al (1985) showed that recall of synthetic speech improves with practice, but is always worse than natural speech whatever the recall task. The differences between performance using natural speech synthetic speech can be related to differences in maintenance rehearsal. Subjects concentrate on the sound of synthetic speech, rather than on the words meaning. This could lead to interference in memory, especially with items presented early in a sequence. Indeed, Waterworth and Holmes (1986) show that synthetic speech shows much lower primacy recall than natural speech.

The major disadvantage of auditory feedback is that it requires a short term store which is easily disrupted, i.e. new auditory information appears to overwrite information previously stored. And takes more time to process than visual information (Robinson and Eberts, 1987). This would be particulary problematic in complex control tasks, where the operator is using a range of commands and data gathered from several sources. For instance, Kidd (1982) has shown that acoustical displays using overlapping categories led to a reduction in subjects ability to select items. If the categories were discrete, performance was good. In the control room, the range of incoming information will be similar to the former condition in this study. Hence, performance will inevitably deteriorate as more information is displayed to the operators.

ii.) Textual Feedback

Verbal feedback can be presented on the screen in one of two ways: either after each word is spoken or after each phrase. In some situations the amount of screen space available for feedback might be limited and so designers may wish to provide single word feedback. Each word recognised will be displayed on the screen, and will overwrite previously displayed words. This will save screen space, but will yield problems similar to those discussed for auditory feedback. Individual word feedback requires the user to remeber which words have been recognised.

In normal operation this might not be too much of a problem, but it could be difficult to correct errors if the user is not sure which words in a command have been mistaken. Obviously, the user could simply repeat the whole command, but this is quite time consuming, and does not always guarantee that the command will be correctly recognised. However, delaying the feedback until a whole phrase has been entered can also cause problems. It is asumed that providing feedback at the end of a long sequence of words is faster than between words (Hapeshi and Jones, 1989). But this introduces problems for the user. The delays between input and feedback could make the correlation of error feedback with the relevant word difficult (Taylor, 1986). Error correction would be time consuming with users having to reenter the whole command.

A solution to the problem of providing verbal feedback on the screen is to provide full phrase feedback, but to display each word as it is spoken. This type of feedback was used by Schurick et al (1985) who found that it provided a faster means of data entry than any of the other types of feedback they considered (see also the telecommand demonstration chapter six). Word for word feedback is used to provide a permanent, whole phrase display. It is not surprising that this type of feedback should the most appropriate for screen based operational feedback. Operators will be familiar with it from their experience with text displays typed in from keyboards.

Research from the field of psycholinguistics can be gathered to support the claim that textual feedback should be presented on the basis of whole phrases, rather than individual words. Rapid Sequential Visual Presentation (RSVP) of text is a commonly used technique in psycholinguistics (Forster, 1970; Masson, 1983). Sentences are presented to subjects one word at a time, at a rate equivalent to normal reading speeds. Masson (1983) showed that subjects who were allowed to 'skim' read a passage were better able to answer questions on it, than subjects who read it by RSVP. As Masson (1983) observes,

> "It appears that the level of comprehension attained during rapid sequential reading may often not involve conscious integration of the elements of text information with themselves or with relevant general knowledge."

Naturally, the type of feedback required will depend on specific applications. Limited screen space might mean that the use of auditory feedback or single word feedback should receive more attention. However, even a small text window could provide the opportuinty to review input data by scrolling. The question is whether any errors need to be corrected as the data is input or whether they can wait until all the data is in.

Feedback could also be used to reduce the potential problems of memory load for medium to large vocabularies. This could be particulary problematic in systems which employ syntax. In these systems, only certain words will be allowed at points in the interaction. If the operator is not sure which words are legal, he might use inappopriate words, leading to errors. It is possible to include a window showing the possible words for that syntax node, however, the user does not always look at this feedback when concentrating on controlling a process (see study seven).

Whilst error correction after all the data is fed into the system may be useful in the fabled 'talkwriter', the more prosaic applications system designers need to consider for actual working environments will require immediate error correction. This requires feedback for each item entered. In a task using ASR to issue commands, the command needs to be checked before it is sent. Therefore, a useful means of feedback will employ whole pharses, displayed on a word for word basis on the operators VDU.

It has been observed in trials using the telecommand demonstration, that a whole phrase display can lead users to send inappropriate commands. This is due to the fact that users can be led to interpret a complete command as a correct command, using the appearance of a whole command as a cue to send it. In other words, although feedback has been presented in operational terms, they are using it as instrumental feedback. The command needs to be checked before it is sent, but such checking requires the user to switch attention between command entering and checking.

iii.) Other types of Operational Feedback in ASR Systems

Some systems provide a facility to check the recognition score of a recognised word. Whilst such information might be useful in learning to use the device, it will probably interfere with actual use. It has been argued in chapter two that recognition accuracy is not a meaningful concept when one considers the use of connected word recognisers, or the use of ASR to perform tasks. Providing the desired words have been recognised correctly, the system will function. Telling the user that a word has been matched to its template with a distance of X will not enhance performance. Indeed, because one can expect users' recogntion accuracy to vary, informing them of their scores could make them concentrate on the use of the ASR device.

Wilpon and Roberts (1986) used a barometer like display to indicate the recogntion accuracy of the words their subjects spoke. They found that it had no effect on performance. In a connected word recogniser running the telecommand demonstration, it was found that providing a similar type of barometer like display to indicate subjects consistency was also ineffective. The main reson for this was that subjects did not tend to look at this display. Unless one is aware of the scale used for the display it is not clear what the information is telling one nor what one needs to do to effect performance.

It could be argued that some of additional feedback ought to be provided to the user if an error occurs. This has been considered in terms of an auditory 'beep'. Whilst this might be beneficial in interactions which are purely auditory, the use of beeps is generally considered intrusive and irritating. The error message could take the form of an additional piece of text on the screen. It could be some global form of feedback, indicating that a word had not been recognised, such as "pardon?". Or it could be more specific, related to the system's syntax, such as "which plant?". However, this will not add any additional information to the user's knowledge that an error has occured. Asking the user whether she intended to say a particular word, is also problematic (see chapter nine).

Conclusions

Inevitably, there will be different feedback requirements for novice and expert users of ASR systems. Novice users constantly monitor their feedback to check their performance. This would suggest that if adequate feedback is not provided, users may be performing two tasks in the interaction with the computer (see chapter eleven). The first would be the task in hand. The second would be a monitoring of the interaction, to check that they are speaking 'properly'. One may even go as far as to say that 'natural language' based interaction would result in unnatural behaviour from the user. By allowing the distinction between user and computer to be blurred, by attempting to disguise the system boundary, designers of 'natural language' based systems force users to draw assumptions from their experience of human communication rather than from computer communication.

The user is involved in a "language game" (Wittgenstein, 1956), in which he supposes certain rules according to which the dialogue is structured. In programming languages these rules are explicit and constrained by the logic within them. Generally they are learnt before the user approaches the computer to use it (as opposed to learn on it). When we turn to the developing natural language interfaces, and to speech in particular, it is conceivable that as technology advances and HCI becomes more "human", the user will have to seek the rules by which the dialogue progresses in much the same way as they do in human communication. But whereas humans often give subliminal cues in terms of speech intensity, posture etc., a computer has only a limited number of options open to provide feedback to the user, and only a limited number of styles of feedback. Perhaps the use of operational feedback as instrumental feedback, found in study three, indicates a change in feedback requirements as a result of experience with the system. However, this must surely relate to the appropriateness of the form of feedback used in the context of the specific task, rather than user experience.

An additional variable which will affect the use of feedback is the accuracy of the device used. If the device is very accurate then it could be argued that action feedback will be sufficient. As long as the device performs the command issued, then additional feedback is redundant. However, it is obvious that current ASR devices are not that accurate, and it is not yet clear whether redundancy may not be desirable in the use of ASR.

Whatever level one uses, feedback in ASR should be immediate. The system should provide an indication that it is recognising words as they are spoken. However, the recognised words should be displayed in a permanent form in order to reduce cognitive processing demands on the user. Feedback should be kept to a minimum, and contain only the words recognised. This will provide immediate, operational feedback to users, which is also displayed in an appropriate context.

Symbolic feedback for ASR

The conclusions presented above, can be useful for purely verbal tasks, such as data entry. But some tasks can be defined as visual, e.g. selecting an area on a touch screen. It could be suggested that visual tasks should be supported by visual feedback (Fitts and Posner, 1967 ; Wickens, 1984). Verbal feedback of a visual action requires translation between processing codes, and operators might find visual feedback preferable. This is because a direct mapping can be assumed between image and object. Therefore, operators attention can be maintained on the primary task.

Chapter eleven discusses the principle of stimulus-response compatibility in system design. Responses are facilitated if they maintain the same processing codes as the stimuli. This means that verbal responses ought to be paired with verbal stimuli, and thus verbla feedback should be used in ASR. If this is desirable then other input media, e.g. touch screens, should be considered as well as ASR. ASR requires the formulation and utterance of verbal commands, whereas a touch screen only requires that an area be selected. This means that tasks need to be described in terms of their visual/ verbal components when they are being assessed for ASR (see chapters three, four and five).

It is suggested that control room systesm should not use auditory feedback. Therefore, feedback should be displayed visually. This could contradict the claims made for ASR, in chapter three, as being useful in allowing 'eyes free' operation and mobility for users. By insisting on visual feedback, we are forcing the user to remain in one place. However, this problem can possibly be circumvented by the use of lasrge LCDs above the mimic diagrams to provide verbal feedback to mobile operators.

Whichever form of presentation is used for ASR feedback in control room systems, it will add to the complexity of system displays. Adding an extra box on VDUs for textual feedback might take up much needed space. The use of verbal feedback could also be held to place an additional task of monitoring recognised words between issuing a command and verifying it. Users can ignore operational feedback and respond to verification cues alone.

Instrumental feedback has been proposed as a means of allowing users to keep their attention on the primary task. A specific command is paired with a specific action by the system. Provided the action is performed, there is no need th check the recognition of the command. But such an approach is prone to difficulties in a system as fraught with error as ASR. Some form of error correction is needed and this requires operational feedback (Rosinski et al., 1980).

An intermediary level of feedback can be proposed between operational and instrumental feedback. It requires the use of symbols rather than words. Control room displays often consist of plant diagrams. When an operator speaks the name of a component, the corresponding symbol on the display could be highlighted. Feedback would be instrumental in that an immediate response would be obtained without any intervening textual material. But it would also be operational in that it conveyed information concerning intended actions, rather than initiating the actual performance of an action.

There would be no need for textual feedback. Rather, the operator would select an object, speak its name, and it would be highlighted. Intuitively, symbols ought to be easier to recognise than words, especially in the context of plant diagrams. In selecting a componenent on a diagram, the operator would tend to look at that component. Symbolic feedback could provide information in the appropriate area of the screen; if nothing appeared then he could assume that an error had occured.

Symbols would also maintain a constant type of information representation onthe diagram. It is a truism that in display design, the operator should not be required to translate between different information codes on the display, i.e. textual and visuospatial. For these reasons, symbolic feedback could be very useful in ASR. However, to date they have received no attention in the field.

Complex spatial relations can be represented more easily graphically than verbally (Gerstendörfer and Rohr,1987). This is supported by Richardson Simon et al (1988), who state,

> "The diagrammatic structuring of information should also reduce the amount of verbal information which is known to produce a higher cognitive load than 'good' diagrams. 'Good' diagrams produce auotmatic control of attention with the help of location objects."

Therefore, it is possible to argue that graphical feedback is potentially superior to verbal feedback, under certain conditions. Unfortunately, there has been little research into what these conditions are. Studies comparing verbal pictorial feedback have generated conflicting results; some studies show superiority for verbal feedback, while others support symbolic feedback. Such studies take the form of either search tasks or reaction time studies. In the former, subjects are required to search for a target against different backgrounds, in the latter, they are required to make a decision concerning the meaning or name of the target as quickly as possible.

Alphanumeric displays have been shown to yield faster search and identification times than graphical symbols (Christ and Corso, 1983; Remington and Williams, 1986). This can be explained by the fact that letters and digits are more familiar to subjects than the geometric symbol used in these studies.

Christ and Corso (1983) demonstrate that after nine months of practice, subjects could identify and search through all sets of symbols with equal speed. This means that subjects are able to learn the meanings of symbols and this will affect their performance. This would suggest that the familiarity of text over symbols could have been an artefact of the experimental situation. If one took symbols and text of similar levels of familiarity, then a more accurate comparisn could performed.

Richardson Simon et al (1988) showed subjects colour patches, geometric shapes, and words for these colours and shapes. The words and pictures were presented in pairs, and subjects had to say whether the stimuli were the same or different. It was found that decisions were the fastest when comparing colour patches with decisions concerning shapes being the next fastest. Comparison of words came next. The slowest decisions occured when subjects were required to make decisions across types. That is, when an attribute (colour or shape) was paired with a word.

It appears that pairing attributes with words requires some form of transaltion. This suggests that there is a difference in the manner in which information is represented for words and for attributes. A dual code theory of independent ,but interconnected cognitive subsystems, for the processing of visuospatial and verbal information (Paivio, 1986) can be employed to explain these results. The issue of information processing codes will be discussed in detail in chapter eleven.

Meyer (1981) provides electrophysiological evidence to demonstrate that colour information is extracted at an earlier stage of processing than shape information. This explains why decisions concerning colours were faster thsan those for shapes, in Richardson Simon et al (1988) and Christ (1975). Presumably, attaching a name to an object will occur after this low level processing. This would explain why it took longer to decide between different information types.

Ergonomic research from the design of road signs appear, at first glance, to be very contradictory. Dewer et al (1977) showed that verbal road signs produced faster recognition times than graphical ones. Walker et al (1965) and Dewer and Swanson (1972) demonstrate that graphical signs are recognised fatser than verbal signs However, in Dewer et al (1977) recognition consisted of naming the sign. In naming verbal signs, subjects simply need to read the word, whereas graphic signs need to be recognised and named.

Road signs need to be proceesed quickly. Their meaning needs to be extracted as soon as the driver sees them. One way of assessing how fast a given sign can be recognised is to present it to subjects using a tachistoscope. This can be used to determine the 'glance intelligibility' of a sign, in other words, how quickly a person can determine what it means. By varying the size of characters in a display it is possible to simulate the effects of distance on perception of road signs. Jacobs et al (1975) showed that graphical signs were less prone to degredation by distance than verbal signs. Thus, verbal signs need not be inherently easier to read than graphical signs.

Pictures are recognised faster if they are primed by a related picture than by the picture's name (Warren and Morton, 1982). This is analogous to the semantic priming of words. Hopkins and Atkinson (1968) found that subjects performance on a task involving the naming of famous faces could be improved if they had learned the appropriate names the day previously. Guenther et al (1980) found that the word 'knife' could be primed by a picture of a knife, and that conversely, a picture could be primed by a word.

Subjects can name words faster than they can name objects (Dewer et al 1976). Further, subjects can classify an object faster than they can calssify a word (Potter and Faulconer, 1975). One of the problems encountered in research comparing the processing of words` and symbols is determining the familiarity of the stimuli to be used. This problem has been ingeniously solved by using the names` and picutes of famous people Hopkins and Atkinson (1968).

Young et al (1986) found that subjects could classify faces of famous people as being actors, politicians etc. faster than they could name them, and that they could speak the name faster than they could classify it. Bruce and Valentine (1985) found similar results to Guenther et al (1980). The recognition of famous faces could be primed either seeing the name associated with the face or by seeing the face itself. Their recognition test required subjects to speak the name of the face. They suggest that two processes are involved in this task. One is to recognise the face as belonging to a specific person, the other is to assign a name to the face.

If the naming of words requires both their classification and pronunciation, then the fact that they are pronounced faster than classified suggests that the pronunciation mechanism gives a faster output than the classification mechnaism. In other words, information form phonological processing is obtained faster than that from semantic processing when the subjects response is articualtory motor. On the other hand, the fact that graphical symbols are calssified faster than they are named suggest that the classification mechanism gives a faster ouptut than the pronunciation mechanism. Marks (1990) states that,

> "Pictures can very rapidly, perhaps automatically, access their semantic representation. Also, it is well established that encoding a picture does not automatically involve activating its label, and conceptual processing may not require phonemic processing."

Therefore, semantic information accrues more rapidly from pictures than from words. Nelson et al (1977) state that picture naming requires greater depth of processing (Craik and Lockhart, 1972) than semantic decisions such as categorization, because naming a picture must occur after some prior conceptual processing (*see also Lupker and Williams, 1989*).

Seidenberg and McClelland (1989) propose a distributed model of word recognition and naming. In this model, naming is the result of a process which constructs an articulatory motor program directly from the phonological code for the word. Lexical decisions, on the other hand, are computed in parallel from phonological and semantic information. Warren and Morton (1982) argue that words can access both a smenatic representation and pronunciation mechanism simultaneously. But the pronunciation mechanism gives a faster output. Objects must access the semantic representation before a name can be assigned to be pronounced.

Study Seven

In considering the use of symbols as feedback in ASR, it is important to consider what the subject will use the feedback for. If symbols are being presented in what is already a crowded display, they coiuld be easily missed or mistaken. It is proposed that when using ASR, feedback has two functions. The first is to indicate that a response has occured to a spoken command, the second is to provide information concerning the validity of that response. Before investigating these points, some studies which investigate symbolic and textual feedback will be reviewed in an effort to develop a predictive model of how feedback will be used in ASR.

Study seven investigates the use of textual vs. symbolic feedback in ascertaining the accuracy of the ASR system's response. This can be regarded as a verbal decision task. The operator speaks a word, and must decide if the system's response is the word spoken. At a basic level, this can be considered analogous to the naming tasks described above. Rather than sepaking a written word, the operator must, in effect, check if a spoken word is correctly written. A direct comparison could be made phonoloigically between spoken and written words, without any intervening semantic decisions required. For symbols, however, subjects must first assign a name to the symbol before comparison. Therefore, it is predicted that textual feedback will be more effective than symbolic feedback for error notification.

Symbols can be defined in terms of their depictive qualties or in terms of assigned meanings. The latter are commonly called symbols and the former pictograms (Sclichcinski, 1977; Barnard and Marcel, 1984). Mead and Modley (1968) distinguish between image related symbols (termed pictograms here), and concept related symbols, such as a right turn arrow. Concept related symbols rely on a strong analogy between symbol and assigned meaning. This suggests that three types of symbol can be used in displays.

Symbols can represent pictograms: objects in defined states, e.g. an open switch. They can be concept related, e.g. in terms of spatial indications. And they can be command words. The latter category is often the most difficult ot design. Symbols cannot easily represent verbs, e.g. open . It is possible to show the action off a paticular verb, e.g. a switch moving from the close to the open position on the telecommand demonstration (chapter six). However, this only shows the result of the command rather than its intention.

Experiments in symbol design tend to take two lines of attack. In one approach, subjects are presented with numerous symbols and asked to rate them for meaningfulness and preference (Guastello et al. 1989) or concreteness and abstractness (Stammers et al., 1989). Although such an approach is a valid method of selecting suitable symbols from a predefined set, it cannot offer an explanation of why symbols are given certain ratings, nor can it explain the processes by which symbols are understood. Consequently, it is proposed that the rating approach will be of little benefit to the investigation of symbolic feedback in ASR.

In the second approach subjects are presented with symbols in the context of simple, perceptual tests, such as visual search or symbol identification. The times taken to find or identify symbols are compared to give an indication of processing speeds for types of symbol. Reaction times in these tests can be used to provide an index of the speed with which subjects can understand information in a given stimuls (Keele,1973). This provides a method of comparing symbols quantatively.

In this study a comparison between subjects' reaction time to different types of visual feedback for ASR will be made. Subjects will speak a word, and feedback will be givewn. It will be either verbal or a type of graphical symbol. Subjects must decide as quickly as possible whether the display matches the word they said. Reaction times will be used to compare performance across the different types of feedback. It is also expected that some learning could occur. This will be less marked for the textual feedback, as it assumed that the words will be familiar to the subjects.

Fourteen students were paid £2 for participation in a study which lasted approximately half an hour. They were told that they were helping to assess the style of feedback to be used for ASR. Subjects were told that the device was speaker independent, so that it would recognise their speech without the need to enrol it. Subjects were assigned to two groups: Group 1 were tested on the textual feedback; Group 2 were tested on the symbolic feedback.

Group 1 were given a list of the twenty two words used in the study, and asked to read them aloud. They were asked to say if any of the words were unfamiliar. This constituted the learning phase for subjects in group 1. After this, they completed the recognition test. The experimenter prompted subjects verbally, i.e. "Please say the 'condenser'", and subjects repeated the word. When the subject said the word, the experimenter pressed a key to call up a word. The words were programmed to appear in a set order, so the experimenter only had to hit the reurn key. When the word appeared, the subject had to decide whether this was the word she had said, and press a 'yes' or a 'no' key. This procedure was repeated five times. At each trial the order in which the subject was prompted to say the words was varied. This allowed some errors to be introduced into the feedback. The rate of error was 25%, which was assumed to mimic poor industrial use.

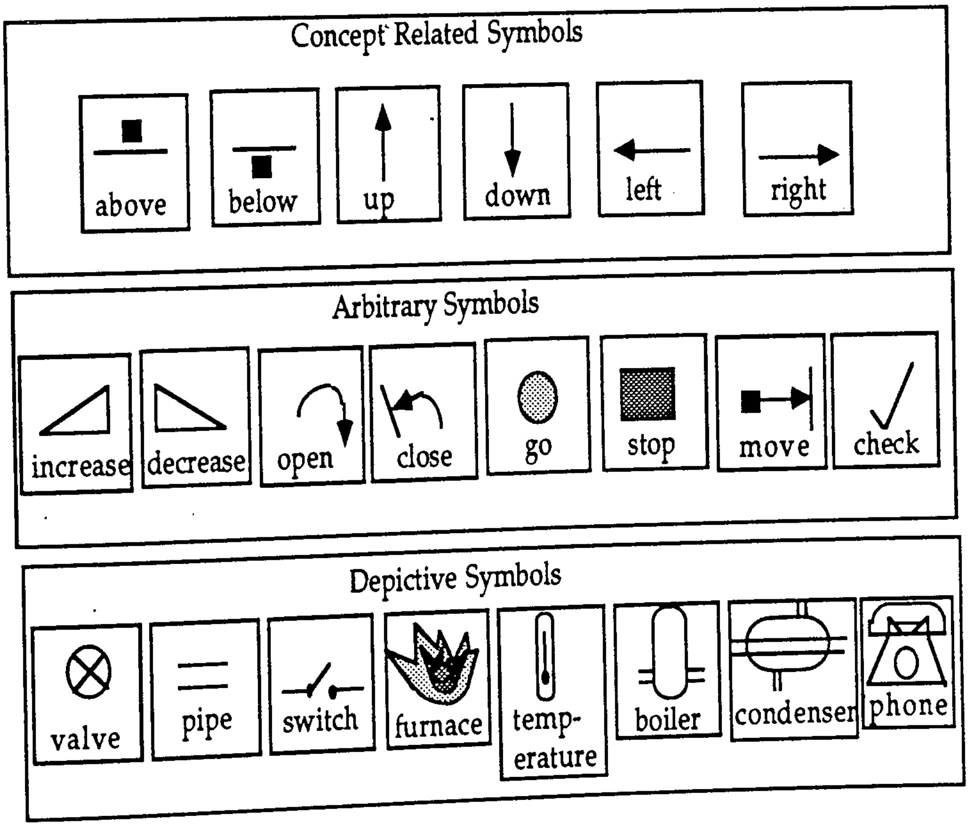


Figure 11.1: Table of Words and Symbols used in Study Seven

Group 2 were shown a table containing all the symbols used in the study. Each word had an associated symbol (see figure 11.1). The experimenter described each of the symbols to the subject. The subject then carried out the learning phase of the study. They were prompted with a word on the screen, below a box containing the symbols. They had to move a cursor onto the correct symbol for that word. This was carried out until all the symbols were correctly identified. The recognition terst was similar to that for group 1, except that instead of words for feedback, group 2 received symbols. Reaction times were measured for the two conditions, and compared in terms of type of item (object, spatial, command) using a three way ANOVA.

Results

Source of Variation	d.f.	Sum of Squares	F.	p.
Feedback	1	20394.148	75.944	0.00001
Error	10	2685.424		
Item	2	724.880	30.022	0.00001
Feedback / Item	2	750.004	31.063	0.00001
Error	20	241.446		
Trial	4	4247.195	14.086	0.00001
Feedback / Trial	4	4885.801	16.204	0.00001
Error	40	3015.091		
Item / Trial	8	735.217	8.276	0.00001
Feedback / Item / Trial	8	820.214	9.232	0.00001
Error	80	888.423		

Figure 11.2: Table of results of ANOVA for Study Seven

There are a number of significant differences across all the measures used in this study. In the comparison of the different types of feedback used, there was a significant difference [F(1,10) = 75.944, p< 0.00001]. Tukey tests revealed that the group receiving textual feedback achieved faster performance than those receiving symbolic feedback, and that this result was constant across all trials.

In terms of the relation between type of feedback and items, there was a significant difference [F(2, 20) = 31.063, p< 0.00001]. A Tukey test revealed that while there are no significant differences between items for textual presentation. There were significant differences between items for symbolic presentation. This is illustrated by figure 11.3 illustrates. In trial 1, there is a significant difference (p<0.01) between all item types. In trial 2, this difference is only between objects and the other item types. There are no significant differences between items in the remaining trials.

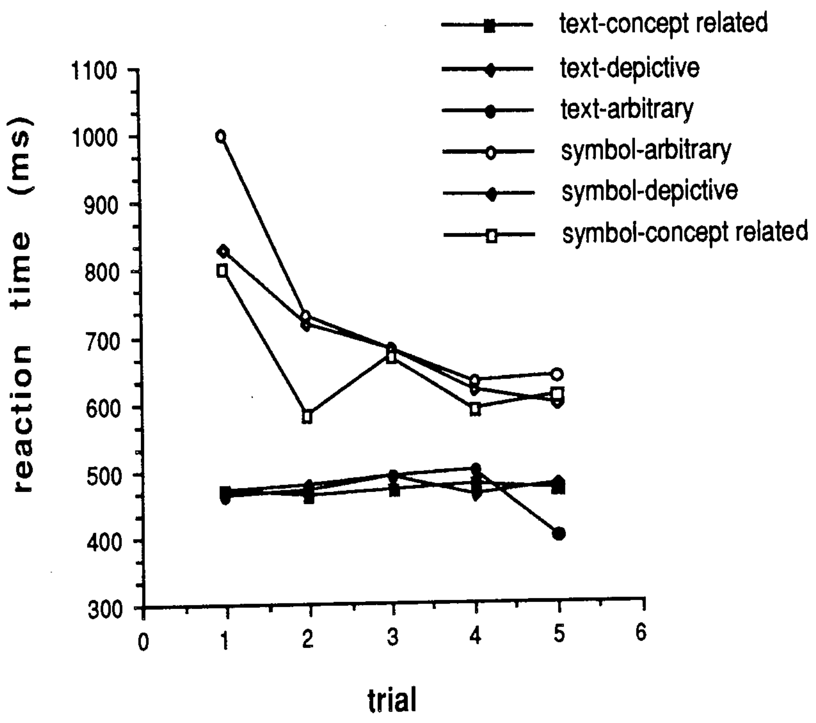


Figure 11.2: Graph of Reaction Time Data for Both Groups

Discussion

Comparison of the mean times for the two groups shows that, as predicted, textual feedback produced a faster reaction time than symbolic. There was no learning effect for textual feedback, performance was constant across all trials and all types. Symbolic feedback showed definitie differences between types of feedback and a learning effect across trails.

The times to make decisions for corect and incorrect feedback were also recorded. It does not seem viable to compare the changes in reaction time across trials for such decisions because a different script was used for each trial. This means that the difference in results over time could be explained by subjects improvement in performance or by the variations in script. However, overall the differences between textual and symbolic feedback can be broken down into reaction times for correct and incorrect feedback, in other words for 'yes' and 'no' decisions. The time taken to respond to correct textual feedback was, on average, 405 ms and the time taken to respond to incorect textual feedback was, on average, 540 ms. This gives a difference of 45 ms between 'yes' and 'no' decisions. The time taken to respond to correct symbolic feedback was, on average, 672 ms and the time taken to respond to incorrect symbolic feedback was, on average, 707 ms. This gives a diffrence of 35 ms. This difference between positive and negative deicions is well established in the literature (Nickerson, 1978; Ratcliff, 1985).

In sum, textual feedback led to faster response than symbolic fedback on a simple verbal decision task. The correction of device recognition errors can be considered as verbal decisions. This can be held to support the suggestion that textual feedback will be most effective for the correction of recognition errors.

Study Eight

Feedback for ASR use can be either for checking the performance of the recogniser, or for checking the performance of the system in terms of a task. We shall call these two types of feedback, Recogniser and Task feedback. For Recogniser feedback, verbal detection is best, as study 1 shows. Subjects had to decide whether the feedback they received was correct or not. With textual feedback this required a pairing of spoken and written words. With symbolic feedback, this required a translation from textual to symbolic codes. For Task feedback, symbolic feedback could be best, because it allows operators to keep their eyes on the selected objects on the screen.

Study eight aims to examine the effectiveness of different types of feedback on the performance of a task. Feedback will be used to indicate recognition of selected objects and actions in a simple process control plant. Obviously, a flashing light or a tone could be provided in response to operator commands. But this would not provide much information about what was being recognised.

In the use of ASR in control room systems, screens are used for monitoring. This requires decisions about the status and position of objects. Symbolic feedback could be proposed to assist in these semantic decisions, as the studies reviewed suggest. Further, the use of a text window could be intrusive on operator performance, requiring a verbal decision to intervene in the performance of the monitoring task.

Ells and Dewer (1979) propose that symbolic road signs are recognised faster than verbal road signs because symbols facilitate faster understanding of the information they contain. But they found that for signs containing small amounts of information, e.g. a single word or symbol, subjects would respond faster to words than symbols. This result is supported by study seven. Steiner and Camacho (1989) found no difference between text and symbols for small amount of information in the display, although they hypothesise that a replication of their study will show text to be recognised faster than symbols. However, as the amount of information in the display increased , response times became faster for symbols than for text. Steiner and Camacho (1989) conclude that for tasks involving semantic decisions, such as diagnosis or question answering, symbols provide a more efficient means of display.

Robinson and Eberts (1987) found that faster response was obtained using diagrams rather than synthetic speech for fault diagnosis tasks. They suggest that,

> "when a system emergency occurs, a practised
> operator will most likely isolate it in terms of a
> spatial reference...rather than as a verbal reference."

Thus, pictorial displays are more compatible with the operators 'mental model' (see chapter one) of the system than textual displays.

While these studies demonstrate the usefulness of diagrams for certain tasks, one cannot use them to argue for the use of symbolic feedback as defined in this chapter. While the use of symbols could conceivably offer the same advantages over verbal feedback as diagrams, or at least support the advantages of diagrams, it is difficult to propose exactly why should be the case. It could be due to the spatial compatibility between symbols and the diagram, or by some undefined relationship between the operator's mental model of the process and symbolic representation in the display.

Study eight compares the use of text and symbols for system feedback (as opposed to ASR Fb) in a process control type task. Symbols will not only allow a direct spatial relationship between objects selected and displayed, but will also assist the development of 'mental models' of the process as subjects progress through trials. A combination of both text and symbols is proposed to present the most effective medium, by giving the best of both text and symbols.

30 subjects (20 men and 10 women), aged between 18 -40 years old , were paid £2 for participating in the half hour studies. Subjects were studying engineering related courses at a technical college. Each of them was assigned to one of the three feedback media in order of participation.

Subjects were required to control a very simple process type simulation on an Acorn Archimedes 440 pc. It was decided that rather than using an ASR device, subjects commands would be mediated by the experimenter. This allowed training time to be minimised, and 'device misrecognitions' to be controlled. Rather than using a "Wizard of Oz" paradigm (see chapter ten), this study involved no subterfuge. Subjects spoke commands to the experimenter, seated beside them. The experimenter then typed in the appropriate command string to the computer. When feedback appeared on the screen, subjects were required to respond as quickly as possible as to the correctness of the feedback, by pressing buttons for 'yes' or 'no'. The performance of subjects was measured by the amount of output they produced, and a prize of £5 was offered for the highest output level. Each subject completed five trials, of five minutes per trial, in one experimental condition.

The three ASR Fb modes were:

Text: subjects received a text 'echo' of their command in a text window below the display (see figure 11.3),
Text appeared as a complete phrase;

Symbol: subjects received symbolic feedback. This took the form of a symbol for 'open' or 'close' appearing next to the appropriate valve;

Both: subjects received a combination of the two types of feedback.

Different feedback was only used to indicate the recognition of commands to open or close valves. All groups received the same feedback for the control of the furnace: a thermometer called be called up to display the temperature of the boiler, and the number of flames lit indicated the temperature of the boiler. It was assumed that opening and closing valves were primarily control actions, whereas controlling the temperature of the furnace and boiler was a combination of supervision, monitoring and control actions.

The layout of the plant used in this experiment is shown in Figure 11.3. This diagram was presented on the screen to subjects in all conditions. The only differences being in the presence or absence of a textual window and symbolic feedback.

Yellow liquid in tanks 1 and 2 passes through the opened valves into the boiler. The furnace can be lit and its temperature rises in proportion to the number of burners lit. If no burners are lit, the temperature drops. Temperature can be checked using the thermometer, which is displayed for a period of three seconds. Once the liquid in the boiler reaches the desired temperature, it turns red. Tanks 3 and 4 contain cold water. Opening valves 7 and 8 allows the cold water to pass over the condenser. If the red liquid is passed through the cooled condenser, it turns green. The analyser (A) will only pass green liquid to the output. The object is to collect as much green liquid as possible.

If the temperature of the boiler rises or falls outside the defined limits, the liquid will return to yellow; if the condenser has no cold water running over it, the liquid will remain red. In either case, the analyser will not pass any liquid to the output. After a preset time has elapsed, the pipe between v2 and v3 will break. This cannot be predicted by the operator as it bears no relation to the ongoing process. The operator must reroute the liquid through v4 and v5.

Final output level was recorded, together with the reaction time data. Also a note was taken of user behaviour. Any mistaken responses by the subject were noted, as misreading errors. Also, note was taken of any feedback ignored by the subject. These errors were defined by the need to point out any feedback to the subject, or by any reaction time in excess of 195 ms, usually these were the same thing.

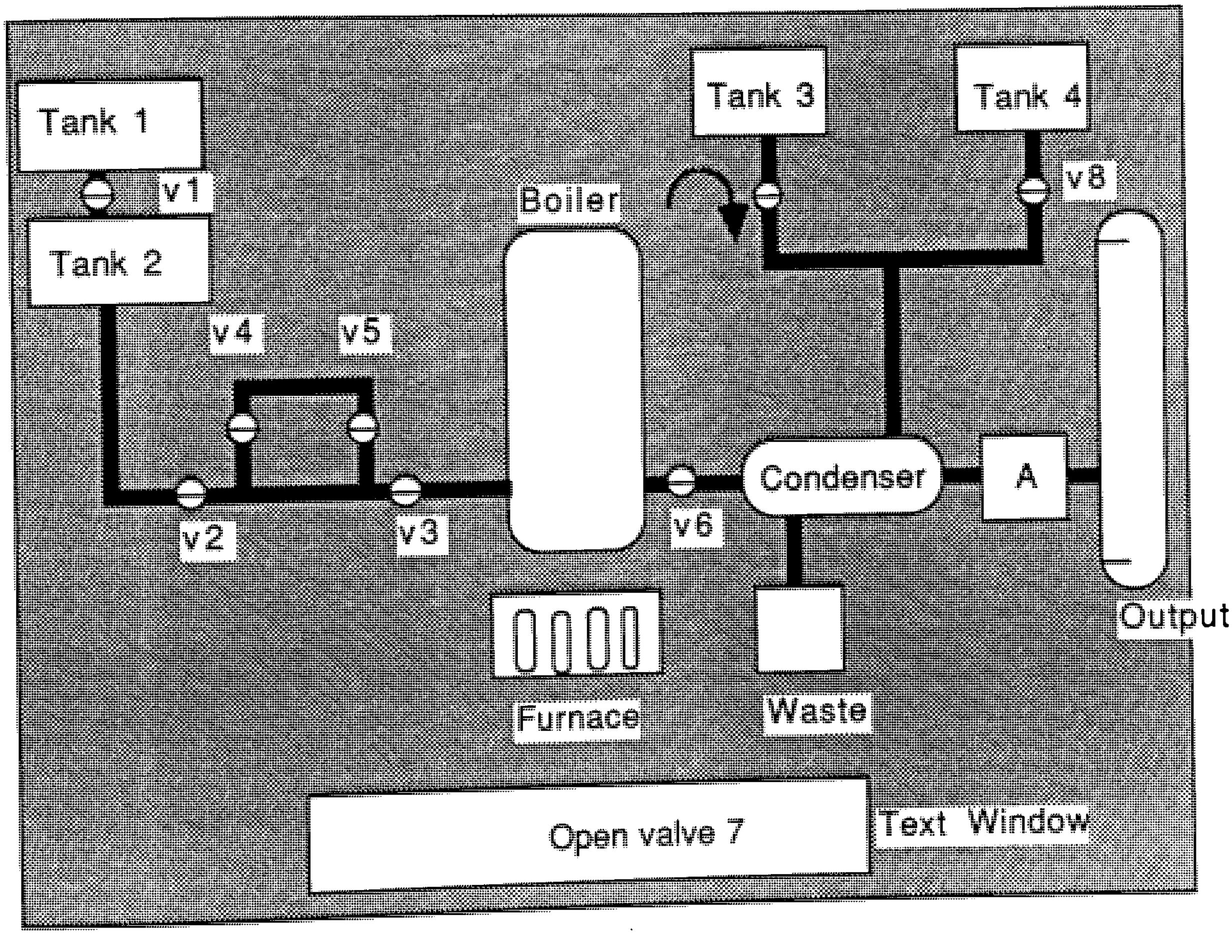


Figure 11.3: Schematic Diagram of Process Plant Display

Results

The results are divided into two sections: one concerning User Performance and the other concerning User Error.

User Performance

User performance, in turn, was measured by the level of output the subjects attained at the end of each trial, and by their reaction time to the feedback.

Output Measures

A two way ANOVA was calculated to compare the effect on output level of type of feedback and trial.

Source of Variation	d.f.	Sum of Squares	F.	p.
Feedback	2	2377.699	1.395	0.27
Error	19	16191.644		
Trial	4	6425.703	5.113	0.001
Feedback / Trial	8	1423.603	0.566	0.82
Error	76	23879.971		

Figure 11.4: Table of Results for ANOVA of output level and feedback

The results showed significant differences in performance: F(4,76)=5.113, p<0.001. A Tukey test was carried out to discover the direction of the difference, and found that while Text and Symbol Groups showed significant increases over trials (p<0.01 for both groups), the Both Group showed no significant change in output score.
This is illustrated by the graph in figure 11.5. There were no differences between Groups at each trial.

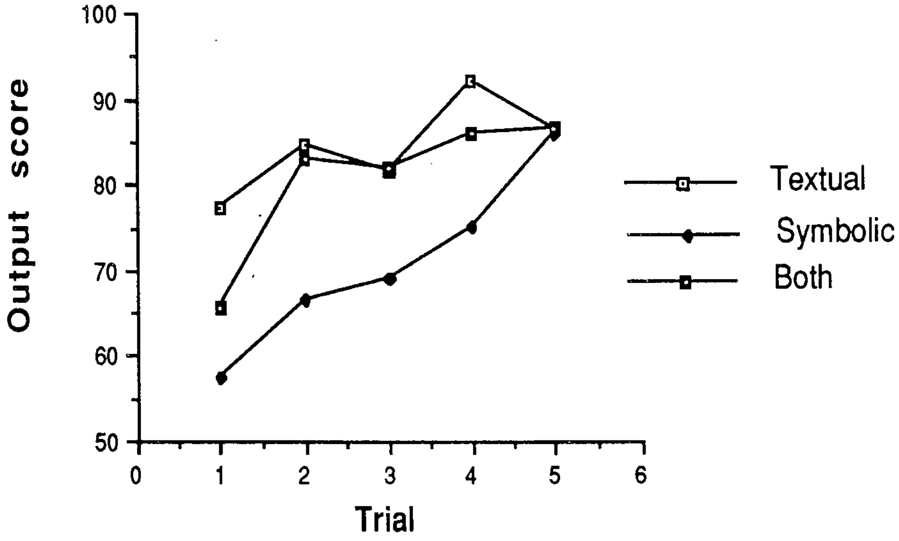


Figure 11.5: Graph of Output Values against Trials for All Groups

Reaction Times

The reaction time data were recorded only for responses to correct feedback. Incorrect feedback was not provided with any consistency, and was used solely as a means of increasing subjects vigilance. Again, a two way ANOVA was calculated for the data, and no significant difference was found between groups. The Symbol Group, however, did show a noticeable improvement in performance over the trials, which supports the result from study seven.

User Error

User error was defined as subjects either ignoring the feedback or subjects responding mistakenly to the feedback. An ANOVA was calculated to compare the number of errors against feedback type. Figure 11.5 shows the results of the ANOVA.

Source of Variation	d.f.	Sum of Squares	F.	p.
Feedback	2	386.568	9.435	0.001
Error	18	368.736		
%errors	1	45.757	1.713	0.27
Feedback / %errors	2	4.849	0.91	0.91
Error	18	480.944		

Figure 11.5: Table of ANOVA for % errors against feedback

Ignoring Feedback

Figure 11.6 shows that text feedback was ignored most frequently. A Tukey test showed that this was significantly more than either the Symbol or Both Group (P<0.01). No difference existed between the other groups.

Misreading Feedback

Misreading errors resulted when subjects accepted false feedback, or rejected correct feedback. Figure 11.7 shows that text produces more misreading errors than the other groups. A Tukey test shows a significant difference between text and the other groups (P<0.01), but not between symbol and both.

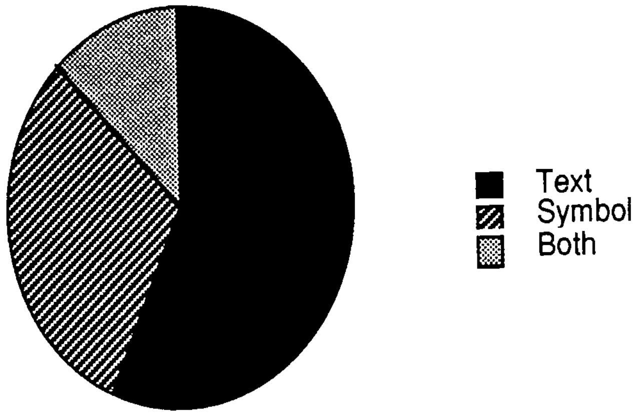


Figure 11.6: Relative Percentage of Feedback Ignored
[Text: 7.35%; Symbol: 4.01%; Both: 1.58%]

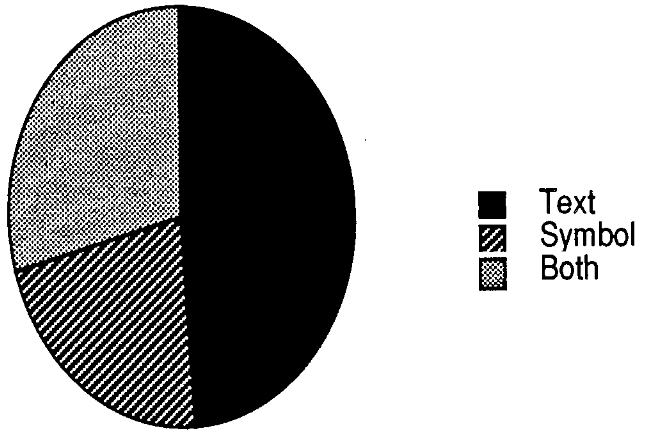


Figure 11.7: Relative Percentage of Feedback Misread
[Text: 9.81%; Symbol: 4.4%; Both: 5.84%]

Discussion

Both the output level and reaction times showed some variation across trials. Text and Symbol Groups showed significant increases in output levels with trials. Thus, practice improved their performance on the overall task. The Both Group showed some increase in output level, but this was not significant. This suggests that subjects receiving a combination of text and symbol feedback, require less practice to control the process.

No significant differences in output level were found between types of feedback in each trial. This suggests that no one type of feedback will provide optimum performance in terms of output level achieved. However, the constant performance of the Both Group points to the use of a combination of text and symbol as being the most efficient for control systems.

Reaction time was primarily included in this study in order to link the results to those found in study seven. It is highly probable that the measure of reaction time in this study contains a more complex range of behaviour than simple reaction time can measure. Ultimately, subjects were required to assess the correctness of the feedback in the context of an ongoing task. Subjects were engaged not only in checking feedback, but also in monitoring and planning.

This said, the range of reaction time in study eight is within the limits found in study seven (600-840 ms in study eight, and 400 -1000 ms in study seven). While textual feedback produced constant reaction time across trials in study seven, there appeared to be a learning effect (albeit not significant in study eight). This is taken to represent the fact that subjects are not only assessing the correctness of the feedback, but also interpreting the feedback in the context of the process.

Further, as in study seven, the performance of subjects receiving symbolic feedback improved over time, suggesting a similar learning effect. This is surprising because the set of symbols used was limited to two which were sufficiently distinct not to be confusable (the data from user errors, below, supports this). Perhaps subjects required time to interpret the symbols, as in study seven. The need for interpretation could be reduced by coding the symbols with different colours (see Richardson Simon et al, 1988). But the use of colour coding leads to a further range of problems for interface design (Christ and Corso, 1983). A combination of text and symbols produces little change in output over time, which we feel, has an implication for the use of feedback by the operator.

The types of user error which were found in this study related either to misreading or ignoring the feedback. study seven suggested that the correction of device errors was a verbal decision task, and required the use of text feedback. But study eight found that for both measures of user error, text gave the worst performance.

This effect is probably due to position of the text window, outside the main display, peripheral to the main task. Consequently, the feedback may either not be seen, or may be not be read properly. The appearance of some text in the window may provoke a response in the operator, before the text is read. This effect was also observed in study three (chapter nine).

The use of symbolic feedback overcomes the problems of positioning, by incorporating the feedback into the display. However, symbols could be prone to misinterpretation, especially given the clear learning effects from these studies. A combination of both text and symbol is recommended because of the low user error produced.

Conclusions

In this chapter we have reviewed several possible types of feedback which an ASR system could use. It was argued that the use of synthetic speech was not appropriate as a means of feedback in control room systems. Therefore, feedback will need to be presented visually. This raises a number of questions concerning the precise form that this feedback should take.

If the feedback is textual, it should be presented after each word has been recognised and in a text window showing the entire command recognised. However, it was proposed that symbolic feedback could be usefully employed for ASR use. This would allow interaction to be based around existing plant displays. The use of a text window led to a significant increase in user errors, but this was probably due to the position of the window outside of the main display, rather than to any information processing code differences. In study three (chapter nine), it was observed that operators often mistakenly 'sent' erroneous messages. This suggests that the 'operational' feedback provided by the text in the window was actually treated as 'instrumental' feedback. Rather than reading the text, operators responded simply to the fact that it appeared.

Study seven showed that verbal decisions, such as error handling, are best supported by textual feedback. This means that device errors will be most effectively dealt with by the operator if feedback is textual. Study eight showed that while textual feedback alone led to more user errors than the other conditions, its use in conjunction with symbolic feedback resulted in the best levels of performance. The use of symbols is argued to provide system feedback related to the process display (and presumably in line with subjects developing 'mental models' of the process). In this way, feedback is integrated into the task. Text provides additional validation of commands, and scope for error handling.

We have proposed a distinction between verbal and semantic decisions, and suggested a relationship between the type of decision and the best form of feedback to support that decision. While study seven showed a relationship between verbal decisions and textual feedback, we are not in a position to draw such a clear conclusion from study eight. While study eight showed better performance in terms of ease of learning and user error resulted from a combination of text and symbol, we are not in a position to offer a substantive theoretical explanation of these results. We suspect that different types of feedback will provide more natural mappings onto the mental models of the process developed by operators. Study eleven, part one (chapter fourteen) shows that the use of a single word adjacent to the valve produces an initially better performance than symbols, but that the difference disappears with two practice trials. This would support the notion that simple decision tasks are best supported by verbal information (see Ells and Dewer, 1979), but not the hypothesised advantage of symbolic feedback in the use of control room systems. The fact that any difference quickly reduces means that, when the operators are familiar with the feedback, then words or symbols could be used. Symbols are recommended as, in general, requiring less screen space than words. Care needs to be taken so that they are easily distinguishable from plant components in the diagram.

A determining factor in the use of feedback is its function in the task, hence different tasks are best supported by different feedback media. The most appropriate medium will be defined by the information processing requirements of the operator and the function of the feedback. So from study seven, the handling of errors is a verbal task, requiring textual feedback, and from study eight, the issuing of commands for specific valves is a spatial task, and requires feedback which bears a clear spatial relationship to the valves in question. Although both words and symbols are of similar efficieny (see study twelve), symbolic feedback is recommended and could be incorporated into existing displays for process operation, with textual feedback in a text window for error handling.

CHAPTER TWELVE

ERROR CORRECTION FOR SPEECH BASED CONTROL ROOM OPERATION

The problem of errors for ASR use have been discussed throughout this thesis. ASR is unlikely to be one hundred percent efficient, and some means of correcting device errors is necessary. This chapter points out that users are also prone to error in using ASR, as study eight illustrated. User error can be dealt with in a number of ways, depending on the type of error.The correction of device errors can performed either by the device or the user. Device based error correction is reviewed, and it is suggested that error correction will always need to resort to the user as final arbiter. User based error correction will inevitably depend on the type of dialogue style they use, and the task they are performing. It is proposed that users be allowed to enter complete commands, and then edit any misrecognised words by repeating them. This removes the need for any error correction dialogue, or any other form of interference with the primary task.

Introduction

Systems which use ASR are unlikely to be 100% accurate. This is not as damning a problem as it first appears. Humans often find that they need to repeat themselves or rephrase their speech in order to be understood. This is a problem with using a medium of communication as imprecise as speech (see chapter one for a full discussion of this point). Users of ASR can expect this aspect of speech use to persist. They must be prepared for the device to make some type of error. Further, in any system which uses a computer to control complex processes, it can be expected that errors will occur. Designers of such systems must regard such errors as inevitable and attempt to design for them. These errors can result from computer malfunction or human error.

Types of Error in ASR Use

ASR is based upon principles which match human speech with stored representations of spoken words (see chapter four). Devices differ in terms of how the representations are derived and how matching is performed. The majority of the techniques employed in commercially available ASR devices attempt to model dynamic speech signals with static representations. This inevitably results in problems of matching, and will lead to various types of recognition error. Techniques for capturing the dynamics of speech, such as Hidden Markov Modelling, have been refined in the past decade and are beginning to be employed commercially. However, such techniques are still subject to recognition error.

Ringle and Bruce (1982) propose that, in human dialogue, failure of the listener to respond appropriately to speakers' words can arise from three factors (under normal circumstances). They define these factors as:

i.) Perceptual Failures - in which words are not clearly perceived, misperceived, or are constantly misinterpreted.

ii.) Lexical Failures - in which the listener perceives a word correctly but fails to interpret it correctly.

iii.) Syntactic Failures - in which all words are correctly perceived and interpreted, but the intended meaning of the utterance is misconstrued.

Of these, only (i.) can be directly applied to ASR use. Both (ii.) and (iii.) require a degree of intelligence on the part of the dialogue partner, which commercial ASR devices have yet to exhibit. Typically, when humans make perceptual errors in speech processing, they are able to call upon their knowledge of the language to correct the error, or they can ask for the speaker to repeat the last word(s) spoken. In situations where such errors could prove critical, for example Air Traffic Control, there have been attempts to reduce the amount of spoken communication human are required to engage in by using other means of information communication (Matthews and Hahn, 1987). Obviously, the latter method of reducing perceptual error is not vialable for ASR use. This means that perceptual errors need to be dealt with either by repeating the misrecognised word(s) or by using some degree of intellignece to resolve errors.

There are three main types of recognition error that an ASR system can produce (Williamson and Curry, 1984). These can be related to the notion of perceptual error proposed by Ringle and Bruce (1982), in that they all involve some form of misperception on the part of the recogniser. The most common type of recognition error is the **substitution** error. This type of error occurs when an incorrect item is substituted for the spoken one. Brown and Vosburgh (1989) found that over 90% of recognition errors, in their experiments, were due to substitution. One can characterise such errors as the misperception of received speech; the user says one word and the deivce 'recognises' another.

Insertion errors occur when spurious noise is recognised as a legal vocabulary item. These account for between 5% and 6% of recognition errors. **Rejection** errors are the least common form of recognition error. They occur when a legal vocabulary item is spoken by the user and the device does not respond, such as would be expected if a problem exists in the communication between user and device. In their studies, Brown and Vosburgh (1989) found that rejection errors accounted for between 2% and 3% of recognition errors. Insertion and rejection errors can be characterised as errors arising because the device could not clearly perceive the words spoken.

While insertion and rejection errors can be minimised by using adequate communication channels, such as a good quality microphone and noise cancelling techniques, substitution errors are harder to define. They can result from the occurrence of similar sounding words in the vocabulary, or from similar templates being created at enrolment. To some extent the vocabulary can be tailored to reduce the number of confusable words, but the vocabulary will inevitably be determined by the tasks one wishes to perform with ASR.

Similar templates are difficult to detect and could result in dissimilar sounding words being confused due to patterns of noise being present at enrolment. This could be reduced by enrolling each word several times, but this could still produce some traces of spurious noise which will lead to confusion. It seems that the only way to deal with such errors, as they cannot at present be designed out of system configuration, is to provide some form of intelligence which can correct them. Such intelligence can be programmed into the device or can be left to the user utilise.

Manufacturers presently claim a recognition accuracy rate of 98%+ for their devices. In the workplace, recognition accuracy varies greatly. We have observed accuracy in the range of 45-90%. This means that recognition errors are highly probable. While errors made by ASR devices in recognising speech have been studied by several researchers, there has been little research into possible causes of user error.

This is somewhat surprising given the current concern in human factors research for design to minimise user error (see Lewis and Norman, 1986). One of the more irksome problems for users of ASR is that recognition errors appear to occur irrespective of user action. Peckham (1986) has noted that a keyboard has a 'standardising' effect on user actions. That is, providing one strikes the correct key, it is unimportant how the key is struck. Using ASR, on the other hand, requires the user to not only speak the correct word, but also to perform the speak the word correctly (within the constraints of the devices matching algorithms). For this reason, the possibility of user error requires investigation. Furthermore, Frankish and Noyes (1990) have shown that, in a task involving entry of strings of digits, while device errors account for approximately two thirds of all errors, approximately one third of errors were due to user error.

User Error

The area of user error in ASR systems has not received serious study. As there are relatively few ASR systems in use, and as these have only been in operation for a short period of time, it is difficult to obtain accident statistics relating to ASR. However, it is possible to draw hypotheses from the general literature on human error (Norman,1981; Reason and Mycielska, 1982; Reason, 1986), which can be tested experimentally.

One can propose user errors will result from an error of intention, such as the incorrect selection of an action, or from errors in execution of the action (Berman, 1986). Reason and Mycielska (1982) define errors of intention as 'mistakes' and errors of execution as 'slips'. Before carrying out an action, a person can be said to formulate an intention to act. This intention need not be clearly defined or articulated. What is important is that it is based on an interpretation of the situation in which the action is to be performed. If the interpretation is incorrect, then the intention will be inappropriate. This will lead to an error of intention, or 'mistake'. If the interpretation is correct, but the action does not proceed as planned, then a 'slip' can be said to have occurred.

Reason (1979) notes that the majority of slips occur in highly practised, overlearnt activities. When a skill is being learnt, extensive use is made of a feedback mode of control to check performance. Once a skill has been sufficiently practised, it becomes automatic (Anderson, 1980). This means that attention is focussed away from the actual performance of the action. Consequently, various types of interference can occur to disrupt the smoothness of the activity. Lewis and Norman (1986) point out that slips are generally less serious than mistakes, and can be more easily corrected. This is because the person making a slip has an idea of what it is he wishes to do. Thus, a slip can be quickly recognised as an aberration from a planned course of action. Mistakes result from a misinterpretation of the situation. This means that people can often not recognise mistakes until they are pointed out or until a problem occurs.

Mistakes can be defined as errors at the planning level of action, specifically in terms of interpreting the situation. Slips can be described as errors at the production level of action. Given this distinction between two basic types of human error, we will deal first with production level errors, slips, as these are the hardest type of error to predict and have, as yet, little empirical support. There has been limited research into the effects of mistakes in ASR use. Methods of reducing user error are considered in the light of results from these papers.

i.)Slips in ASR use

In terms of using ASR slips can be related to the production of spoken commands. The simplest type of slip would occur when vocabulary items are mispronounced due to users introducing spurious noise, such as yawning, into their speech. Users could introduce overlong pauses into their commands, as a result of being distracted by another task. There is an extensive literature concerned with speech errors known as 'slips of the tongue' (Fromkin, 1980), but it is difficult to propose why such slips occur, or how to predict them. Even though it is not possible to predict where slips in speaking will occur, it is important to provide the user with some means of correcting system errors that these slips cause. Error correction is discussed in section three.

Slips can also occur in the interaction between user and computer. Frankish and Noyes (1990) have found that if users receive feedback visually, they are prone to errors in detecting misrecognitions. That is, users do not notice the misrecognitions. This is not the case when feedback is auditory. Ito et al (1989) found that when auditory feedback was provided, users would wait until their speech had been echoed before speaking the next word. For isolated word recognition devices, at least, forcing users to wait until the device has correctly recognised a word will reduce the likelihood of user error. Thomas and Rosson (1984) found that, given the opportunity, users prefer to interrupt synthesised speech messages. Presumably this allows the user to control the rate at which they receive the message, to make processing easier.

Study eight showed that if feedback was presented in a text window, below a display of a process plant they were controlling, users ignored around 7% of feedback. If the feedback was symbolic and incorporated into the display, users only ignored 4%. These results suggest that if feedback of recogniser performance is incorporated into the users' primary task, they will be less likely to make slips in error detection, than if feedback monitoring constitutes a task in itself.

ii.) Mistakes in ASR use

Mistakes could result from a number of factors in ASR. The user might attempt to use an illegal word to issue a command. This could be due to the user being more familiar with a synonym of the command word than the word employed. For this reason, vocabularies need to be designed which users find easy to use and remember.

Poock (1980) has noted that function keyboards provide memory cues for users, in the labelling or coding of the keys. ASR does not normally offer such cues. Legal vocabulary options could be displayed to the user, but this may require screen space which is not available. If the vocabulary was designed to be not only task specific, but also "habitable" (Watt, 1968), users would know which words were required for which operations. This suggests that efficient vocabulary and dialogue design could reduce such mistakes.

Another likely source of user errors is the fact that users can often have difficulties when faced with the restrictions imposed on their style of speech by isolated word recognition devices: users try to speak too fast for the device. However, Brown and Vosburgh (1989) examined the occurrence of 'segmentation errors' in the use of an isolated word ASR device. Segmentation errors arose when users either spoke too quickly for the device, coarticulated words, or spoke too slowly, introducing pauses into words. Of all the recognition errors they recorded, Brown and Vosburgh (1989) found that segmentation errors only accounted for between 0.3% and 0.8%.

This shows that users are capable of adapting to isolated word ASR devices, given adequate training and practice. With the increasing availability of connected word ASR devices on the market, one might question the utility of this observation. However, it supports the observation that users are capable of adapting efficiently to ASR use (see chapter ten), even when the ASR device changes the form of interaction (Zoltan Ford, 1984).

Given that ASR is not 100% efficient, it is necessary to provide users with feedback concerning the performance of the recogniser, and a means of correcting errors. Users can also make mistakes in their use of feedback from the device and in correcting recognition errors.

Study eight showed that, in addition to ignoring feedback, subjects were also prone to misreading feedback. If feedback was provided in a text window, adjacent to the process they were controlling, subjects misread almost 10% of feedback. "Misreading" was defined as the inappropriate response to feedback, in terms of confirmation of commands. Symbolic feedback, incorporated in the display next to the valve to which it referred, was misread on only 4% of occasions. Again, this suggests that incorporating feedback into the primary task will reduce the likelihood of user error.

Automatic Error Correction

Before an error can be corrected, it must be detected. Errors can be detected if the input speech violates some set of rules. These rules can either be held in the device or by the user. Obviously, the latter type is the easiest to implement and will be discussed below.

Error detection and correction rules can be held in the device in two ways. They can either form part of the recognition process or can be applied after recognition. Rules which form part of the recognition process were briefly discussed in chapter four. They can be said to be essential to intelligent ASR. Such rules ought to enable ASR devices to reduce errors to a negligible level, as the incoming speech is assessed for meaning and relevance to the ongoing task. Such implementations are still in the early stages of development, despite some of the successes mentioned in chapter four. Green et al. (1983) point out that the only internal source of error detection available to most ASR devices is the goodness of fit between template and word. If the goodness of fit is below a certain threshold, the word is rejected. Thus, erroneous words can be rejected. This technique works on an all or nothing basis. If the word is sufficiently well recognised, it will have a good recognition score. But it is possible for words with high scores to be wrong. Therefore, some form of error correction is required in order to check and correct input speech.

Error detection and correction strategies which can be applied after recognition appear easier to design. The recognised word can be assessed in terms of the context of the rest of the command. This assessment can take a number of forms, depending on the definition of context. The simplest definition of context relies on the syntax of the incoming command, e.g. a sentence can be said to be made up of a noun phrase and a verb phrase, each with definable constituents. A set of syntactic rules can be used to provide information for a simple parsing routine. Tomita (1986) used such a set of rules in a context free grammar to provide information for a left to right parser. The incoming words form a string as they are recognised. This string can be parsed until a syntactically well formed string is produced. Such an approach is of limited value because it assumes that all the words spoken are have a recognised word associated with them. This means that the routine cannot handle insertion errors. Also, because the routine does not use any operational information, it is possible to create syntactically well formed, but meaningless strings.

Many commercially available ASR devices offer a second choice word form the matching process. This is the template which has the next closest match to the word recognised. This can be exploited to provide a means of error correction. It is argued that that substitution errors arise from similar sounding words. Thus, if the recognised word is wrong, it is reasonable to assume that the second choice word will be the required word.

Spine et al. (1983) initially allowed subjects to select the second choice word to substitute for the recognised word when an error occurred. Such an approach proved popular and was reasonably effective. It was reasoned that if sufficient syntactic and semantic information was incorporated into the software, then a simple parsing routine could be used to correct errors.

The routine designed by Spine et al. (1983) comprised three stages. If the recognised word did not conform to the rules, then the second choice word was tested. Any second choice words which did not conform to the rules were deleted. The number of correct messages transmitted rose from 63.5% to 82.9% when this form of error correction was used.

Although the approach appears attractive, it has a number of problems. A second choice word might have a similar template to the recognised word, but there is no guarantee that the recognised word is correct. If some form of insertion error occurred then the use of second choice words for error correction is of little use. Rejection errors cannot be handled because the device needs some word in order to carry out the error correction. Some words can be distinct enough not to have a very close second choice, what happens when these words are misrecognised? The second choice might be made on a part of the template rather than the whole word. This means that erroneous words could be selected. Second choice words are only generated on recognition. Therefore, this error correction technique can only work on a single pass. Deleting the word requires the user to repeat it. Whilst this could help to reduce the amount of repetition required of the user, it questions the name of automatic error correction. The simplest parts of the business of error correction are given to the device, leaving the user to monitor the progress of both the recognition and the error correction processes.

Error correction which allows the device to guess at missing words is a more attractive proposition. The error correction decisions could then be said to be truly automatic. Hayes et al. (1986) propose using a caseframe analysis to drive their parsing algorithm. This enables the parser to anchor its interpretation on the most significant parts of the input, called "head concepts", e.g. nouns and verbs. From these "head concepts", any gaps in the message can be filled in with propositions and pronomials which are phonetically similar to the recognised words. As the authors point out, whilst most of the "head concepts" in their vocabulary were multisyllabic and acoustically distinct, many were difficult to recognise. This suggest that the sue of "head concepts" is a clever but misguided approach to error correction. A more practical solution would be to drive the error correction process using the most easily recognised words, regardless of their importance in the sentence. This is the approach being developed by Dreizin (1987).

The 'sieve' method of error handling, as Dreizin's (1987) approach is called, begins by overriding two assumptions central to parsing natural language. Rather than dealing with the input in a left to right manner, words which satisfy global constraints are used as the basis for the initial part of the error handling process. Errors are not assumed to be special cases but are the norm, with well formed strings being special cases. Such assumptions make sense for the recognition of strings in ASR.

Error recovery involves replacing erroneous words in the string with phonetically similar words which do not not violate grammatical constraints. These constraints are represented by global concepts which are represented in a dictionary. Each word in the vocabulary has a dictionary entry listing the constraints which apply to it. The constraints are combined to produce a 'sieve' which sifts all the first choice words recognised, and supply suggested replacement words until a well formed string is generated.The following example will illustrate this process. The vocabulary used relates to the domain of artillery fire control. Suppose the user issued the command:

"COMMAND POST IN TREES OVER"

and the device recognised this command as:

"COMMAND FOUR IN THREE OVER"

The error recovery process would begin with the dictionary entry for the word COMMAND, shown below.

word:	COMMAND
properties:	MAINTARGET UNCOUNTABLE CCEWORD COMMAND
properties forbidden to the right:	ADDINFO MAINTARGET DIGIT
properties forbidden to the left:	DIGIT . MAINTARGET [condition:(not(property (F,POST) or property (F, VEHICLE))] ARMINFANTRY ENEMY ATTACKING ASSEMBLING
obligatory property adjacent to the right:	POST

From this dictionary definition, the word COMMAND cannot be followed by a digit. As there are two digits to the right of COMMAND, this represents two constraint violations. According to their dictionary entries (not shown here), DIGITs cannot have MAINTARGET words to the left of them, and must have at least one MAINTARGET to the right. FOUR is preceded by a MAINTARGET, and is not followed by one, causing two further violations. THREE is preceded by a MAINTARGET and ADDINFO (i.e. IN), both of which are violations. Finally, according to its dictionary entry, ADDINFO cannot be followed by a DIGIT, so THREE after IN is a violation. Thus, seven violations occur in all.

Error correction begins with the words which are most consistently misrecognised for a given speaker. These words are defined by previous performance. A phonetically similar word is substituted for these words when violations occur. TREES is substituted for THREE, and the number of violations is immediately reduced from seven to three. The word COMMAND must be followed by the word POST. Substituting POST for FOUR reduces the number of violations to zero.

Obviously care must be taken over the design of the grammar, but current performance results are very promising. A ten word string can be analysed in approximately two seconds. The number of correctly transmitted messages rose form 55% to 77% when error correction was used.

A problem which is common to all automatic error correction algorithms is that syntactically or semantically well formed sentences can still be transmitted. The sieve method described above goes a long way to reducing this problem, but the sentences generated can still be wrong in terms of what the user actually intended to say. In such instances ,only the user can check the validity of the recognised message. Therefore, it is advisable that some form of user controlled error correction be employed.

User controlled error correction

User controlled error correction techniques can be usefully classified using a set of headings proposed by Hapeshi and Jones (1989). These headings relate to the amount of data which is input before error correction can take place. Error correction can occur after each individual item or word has been entered. It can take place after a line of data or a phrase of a command has been entered. It can occur after a whole command, or after a paragraph, or after the complete document.

Casali et al (1988) found that subjects were able to recognise the higher levels of ASR device accuracy (from a range of 75%, 87.5% and 100%) and preferred these to lower ones. Error correction was carried out by spelling misrecognised words, which users found frustrating. The frustration did not appear to arise from the fact that subjects had to spell words. A further variable in the study was "available vocabulary". This related to the number of words the subject could use. If a word was not in the 'available vocabulary', the subject was required to spell it. This did not effect acceptability ratings. Therefore, the fact that error correction interrupted the flow of the task was found to be frustrating. This finding can be considered in terms of the notion of locus of control (Rotter, 1971).

Locus of control describes the extent to which an individual can feel in control of his behaviour in specific environments. An external locus of control emphasises factors or events in the environment as being responsible for an individual's behaviour. An internal locus of control emphasises the individuals responsibility and control. As Simes and Sirsky (1986) point out,

> "Computer controlled conversations often frustrate people who have a strong internal locus of control, and reinforce the perception of helplessness for people with a strong external locus of control."

This frustration generally results from the power relation of the dialogue being too far in favour of the computer. Gaines (1981) proposes that the user should dominate the interaction. Errors lead to an instability in the dialogue, and to uncertainty. This in turn can lead to frustration, with users not sure what is happening or what to do next. Consequently, a consistent balance of control is necessary. By requiring the user to perform the simple, but intrusive task of spelling misrecognised words, the power balance has shifted causing a shift in locus of control. If the user could either say "no" or repeat the word, the he would remain in control.

The simplest form of error correction would require the user to say "yes" or "no" after each item has been recognised and displayed. This will produce a fail safe system, but will be extremely time consuming to use. Little and Joost (1984) suggest that the user need not respond to correctly recognised words, and simply respond "no" or "delete" to misrecognitions.

This type of error detection dialogue can be extended to cover whole commands. Martin and Welch (1980) propose that a buffer could be used to store words as they are spoken. Verbal commands could then be used to edit the information in this buffer before the command is sent. For example, the user could say "o.k." to verify the whole of the command in the buffer, or they could "erase" individual words, or "cancel" the entire command. This approach was tested by

Spine et al. (1983) and Schurick et al. (1985).

Spine et al. (1983) found that where subjects had the option to correct errors on an individual item or whole command basis, they tended to prefer individual item error correction. Their studies showed that individual item error correction was used on average 114 times, compared to an average of 6 times for whole command error correction. Similar results were obtained by Schurick et al. (1985).

In their studies, Schurick et al. (1985) found that individual item error correction was used 48% of the time, and whole command error correction was used only 12% of the time. It is interesting to note that if feedback was presented in the auditory mode, subjects preferred to correct the whole command rather than individual items. This can be accounted for by the memory load imposed by auditory feedback, as discussed in chapter eleven.

These approaches to error correction employ a model based on a very limited text editor. Martin and Welch (1980) point out that an obvious problem in the use of verbal error correction commands is that these commands might themselves be misrecognised, and suggest that,

> "This problem must be minimised either by an inherently low error rate or by choosing the correction commands to be as phonetically different from the other words in the vocabulary as possible."

These are potential solutions, but it is suggested that the text editor model suggested here is not appropriate for industrial speech based interaction with machines. If the device misrecognises a word, the user needs to issue a command word before he can repeat the misrecognised word. It seems far more sensible merely to repeat the misrecognised word than to use command words. The likelihood of the command words themselves being misrecognised makes it difficult to argue for an inherent advantage in using such an approach.

The error correction dialogue imposes an intervening task between the user and the primary task. As the word appears on the screen, the user is required to make the following decisions:

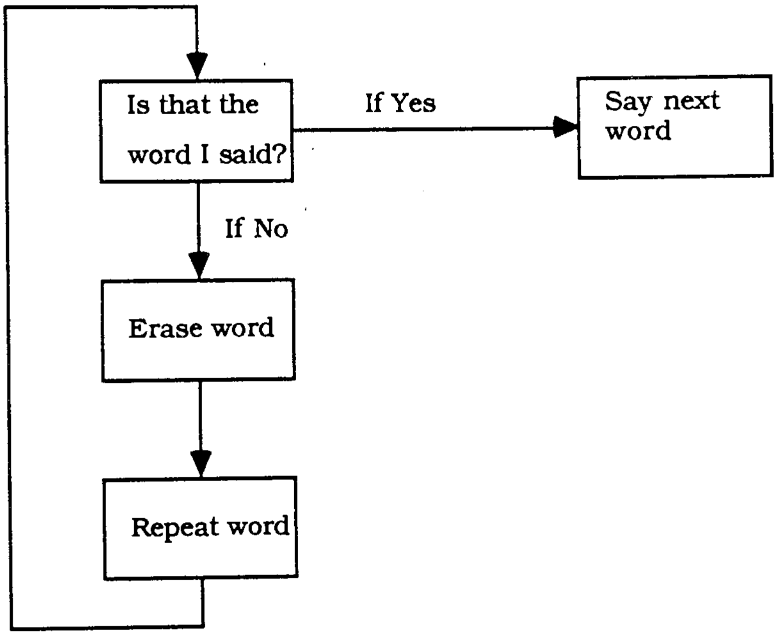


Figure 12.1: Error Correction Dialogue

A simple game was devised (by D.M. Usher) which used voice commands to move a ball through a maze. The size and velocity of the ball could be altered, together with its direction of travel. Error correction employed a dialogue based on the proposed word. The device used a threshold to reject words. If a word was just below the threshold, the user was prompted with, "Did you say x ?". She had to respond either "yes" or "no". If the user responded "yes", then the word was accepted and the action performed. If the user responded "no", then the user could repeat the word or update its template. Such an approach appeared intuitively feasible, and was sophisticated enough to offer quite extensive user control. However, when prompted "Did you say x?", users often tended to repeat the misrecognised word. The prompt was treated as evidence of misrecognition rather than as a cue for information. The tendency of users to repeat misrecognised words has been noted extensively in the literature. Study five (chapter ten) found that subjects not only repeated misrecognised commands. but also spoke in a louder tone of voice. This will inevitably lead to a decrease in performance. The louder a person speaks, the more their speech will drift from the recorded template. Therefore, words should be repeated in the same tone of voice as they were initially spoken. Asking users to carry out an error correction dialogue might appear more 'human', but is in fact less normal than allowing them to repeat a word, as if the device had a hearing problem (see study five, chapter ten).

The studies presented show that users prefer immediate error correction of individual words. They also suggest that, in agreement with Schurick et al. (1985),

> "[subjects] selected correction commands they
> thought would be most suitable in a given situation."

It is suggested that error correction dialogues are not necessary in industrial based ASR. Most control room applications of ASR will centre around the use of ASR to issue commands or enter data. In such situations, recognised speech needs to be corrected before it is sent to the host computer. Thus, error correction needs to be immediate and as simple as possible so as not to interrupt the smooth performance of the primary task. It could be argued that as more text is entered, then more text editing facilities will be required. However, it would be very cumbersome to shift a cursor using spoken commands. Consequently, it is suggested that in such situations, text be edited manually. Chapter seven discusses some of the issues of using ASR in conjunction with other input media.

The telecommand demonstration (described in chapter nine) offered three types of error correction for a whole text display. As each word was entered, it filled a predefined slot. When an error occurred, subjects could either repeat the misrecognised word or complete the command. Error correction could be conducted either using individual items or whole commands. Consequently, the types of error correction were as follows:

> i). immediate, individual item error correction, i.e. each item entered and, if necessary, corrected before the next item is entered.
>
> ii). individual item error correction after the whole command had been entered, i.e. subjects spoke the whole command and then repeated any misrecognised words until the desired command was constructed.

iii). whole command error correction, i.e. subjects simply repeated the whole command if any errors occurred.

Unfortunately, it is not possible to offer any comparative figures of the use of these types of error correction. In practice, subjects always used the second method.

The use of repeating commands to correct errors might appear overly simplistic. However, it is argued that this approach is more natural that error correction dialogues. In these trials, by waiting for the whole command to be entered before correcting errors, subjects were not required to interrupt their primary task of issuing telecommands. A command is issued. Then it is checked and corrected. Then it is sent.

If words are consistently misrecognised then repeating them will only lead to an increase in frustration. This will probably effect the way the user speaks, causing the speech to drift even further from the templates. The only method of correcting such persistent errors is to update the templates. Chapter thirteen discusses some approaches to this problem.

Conclusions

It is proposed that for industrial speech based interaction with machines, user controlled error correction is essential. Such error correction needs to be provided out at logical points in the interaction. In the studies reported, this was at the end of each command entered. Other situations could have different requirements. However, the unit of data entered will determine when error correction can be applied, so as not to interrupt the performance of the users' primary tasks.

It is further proposed that an error correction dialogue might appear intuitively feasible, but is inappropriate. Command words might be misrecognised. The model of text editing for error correction in ASR is mistaken, and a more natural means of error correction will allow the user to repeat misrecognised words.

The templates of consistently misrecognised words need to be updated. Obviously the type of error correction employed will depend on the application, the type of ASR device used, and the environment in which the device is to be used. However, this discussion is aimed at the use of ASR in control rooms and provide a set of recommendations which are considered appropriate to, and implementable in, such an environment.

CHAPTER THIRTEEN

DEVICE AND USER TRAINING

It has been suggested that control room systems use speaker dependent ASR. This means that, initially, operators will have to enrol the device before they can use it. This chapter presents a discussion of how best to carry out enrolment. Also, operators need to be trained before they use ASR. This is considered in a study, which shows that a demonstration by an experienced user of ASR produces better performance than verbal instruction. This result is considered in terms of implicit/explicit knowledge.

Enrolment

The majority of commercially available ASR devices are speaker dependent. This means that before a person can use a device, it must be trained to recognise the vocabulary items as that person speaks them. Many writers use different terms to describe the process of creating templates of the user's voice in an ASR system. Some examples include: voice template creation, voice training, system training, speech sampling, data collection. I prefer the word enrolment, because it allows flexible use (i.e. to give the forms enrolling, enrolment, to enrol etc.) and is relatively unambiguous in the context of ASR literature. Any variation on the word 'training' could conceivably be confused with user training.

The area of enrolment has received much attention, and is well researched. Consequently it is possible to design an enrolment procedure which can be applied to the majority of ASR systems. Conversely, the issues surrounding user training have not been studied in any depth. Manufacturers tend to provide prospective users with simple lists of instructions. Study ten investigates the use of demonstrations by an experienced user as an alternative to the use of instructions.

Separate vs. composite templates

Speech is inherently variable (Lieberman and Blumstein, 1988). This means that each utterance of a particular word will be slightly different. This difference is exacerbated when the speaker is tired, ill, or under stress. Commercially available ASR devices employ template matching principles, with spoken examples of words stored as templates. Chapter four showed that templates capture broad acoustic properties of the speech signal, rather than linguistic information. This analysis is beneficial in that it removes some of the possible sources of variation. It is essential that the effects of variation in speech are overcome, and that templates are as representative of the users speech as possible.

One way in which enrolment can be made to capture these representative templates, is to provide several examples of each word. ASR devices either store each enrolment of a particular word as a separate template, or they combine enrolments into single, composite templates. In the separate template method, two parameters place a limit on the number of items that can be stored. The first is the available memory space, and the second is the acceptable time for recognition. Obviously the more items stored, the longer the time to compare the incoming word with each item.

A possible advantage of storing separate templates for each utterance of a particular vocabulary item is that each template might contain slight variations of the utterances. Taken together, these templates could capture the range of variation in several utterances of a single item. Whilst separate templates might help to capture such variation, there is the possibility that two templates corresponding to the utterance of different vocabulary items will become confused. Such confusion is, of course, inevitable with any template matching technique; a similarity in a small part of the templates could provide sufficient information for confusion. However, it seems plausible that the more recordings one makes, the wider the range of possible confusions, e.g. two templates of acoustically dissimilar words might contain sufficient breath noise to provide similar templates. Often there is no way of knowing whether such an error is likely until the templates are compared. Even then it is difficult to know what has caused the confusion, unless one can look at the templates.

The Marconi SR128 used in this research, allows users to investigate templates as hexadecimal codes. D.M. Usher, at the CEGB, has written software which converts these codes into colour diagrams. With some practice, it is possible to interpret these data to such an extent as to be able to note points of similarity between templates. But much of this interpretation relies on intuition and guesswork. Further, it is time consuming and one which is unlikely to be automated in the near future (cf. the discussion of expert spectrogram reading in the chapter four).

A further problem with single templates concerns inter item confusions. Whilst single templates might capture some of the variation in the pronunciations of individual items, there is the possibility that templates for different items might become confused. One way around the problem of individual template confusion would be to provide a template counter in the software. This would count the number of templates selected for each vocabulary item. The item with the most templates selected would then be returned as the recognised word. One could develop this idea to ensure that templates that were continuously problematic, i.e. confused with other templates or not selected at all, could be erased and re enrolled. Such an idea is a simple version of template adaptation discussed below.

One of the reasons that composite templates were introduced was to reduce the amount of memory which needed to be reserved for storing templates. Rather than having four templates representing each pass, i.e. utterance, of one vocabulary item, the ASR system produces an average of the four passes, and stores them as a single, composite template. Williges and Williges (1982) found that if too much information is included in such a template, it will become at best 'messy', at worst meaningless. They suggest that anything over nine passes per vocabulary item will produce unmanageably noisy templates, and a subsequent decrease in performance.

The use of composite or single templates will depend on the device used. There are advantages and disadvantages for each. However, the composite template approach seems best suited to the matching process in ASR. Rather than attempting to match against several examples, the device uses a single average template for each word. The number of passes required to create such an average template is discussed below.

Number of passes required for enrolment

One can suggest a lower limit of one pass and an upper limit of ten passes per vocabulary item, whichever form of storage is used. For separate templates, this figure will relate to the amount of available memory space. Early commercial devices required ten passes per item, but had vocabularies of only ten to twenty words. Manufacturers now claim their devices will work adequately with single pass enrolment. In study three, it was found that a single pass enrolment was sufficient if the device was to be used immediately after enrolment. Bearing in mind the variability of speech, it seems more sensible to provide a number of passes per item.

Mutschler (1982) compared one, two, and three passes per item in a vocabulary using German digits 0-9, on a Speechlab 20A device. He found that percentage recognition rate, averaged across twelve speakers, increased with the number of passes used. One pass yielded 79%; two pass, 87%; and three pass, 94%. No significant improvement was found for three, four, and five passes.

Another study, reported by Mutschler (1982), was carried out by Batchellor (1981). Recognition accuracy for a fifty word vocabulary on a Threshold T600 device was compared for different length enrolments. A three pass enrolment gave 97.3% Recognition accuracy; five pass gave 99%. Seven pass enrolment gave no further improvement. Thus, it would seem that enrolment of between three to five passes ought to capture sufficient of the variation of speech to yield acceptable performance.

If the vocabulary one is using is quite large then enrolment will be time consuming and boring for the user. Until there are inexpensive and usable speaker independent devices, it seems one has little option. One could tailor the length of the enrolment schedule to fit the attention span of the user, perhaps requiring a run through of the whole vocabulary once a day for three days. This is intuitively reasonable because day to day variations might be captured. Further, levels of arousal are known to vary throughout the day (Wever, 1988). Such variation can effect the voice, with low arousal producing quieter, more tired speech. Connolly (1979) reports that he uses two sets of templates, one for the morning, and the other for the afternoon when his voice is tired from speaking all day.

Alternatively, one could enrol each vocabulary item once only. During the course of use, some vocabulary items seem to be recognised consistently well, while others create problems. The problematic words could be re enrolled as required. This would lead to several examples of each word being stored, but only when required. Thus, enrolment time would be reduced, but sufficient examples of each word ought to be stored.

Presentation of items for enrolment

It is common for users to be presented with a list of the vocabulary items and asked to speak them in order. The user will either repeat each word several times before moving on to the next one, or will repeat the list several times. Both approaches can be seen to be undesirable. Linguists have long been aware of the occurrence of 'list intonation', i.e. the pattern of falling and rising tone we adopt when reading a list of words. Such a style of speaking is very different to that used to issue commands. It is this latter style we need to capture during enrolment of industrial devices.

Research from telecommunications (Kryter, 1972) suggests that lists are best transmitted with some variation in word order to prevent speakers, or listeners, from slipping into 'automatic pilot' when running through the lists, and mistaking spoken words for expected ones.

Poock (1980) found that varying word order on enrolment passes helped to increase system performance. Varying word order can be assumed to maintain a sufficient level of arousal to check the word they are saying. This could reduce the monotony of the task. The introduction of such variation could also capture some of the variability of speech , with words being presented in differing linguistic contexts, i.e. with different words preceding them. This would remove the burden of attempting to remember how one enrolled a word in order to duplicate it during use.

There are obvious problems with varying word order when one is dealing with preset presentation orders. The user types in the words to be enrolled. These will provide labels for the recorded templates. The labelling is achieved by the matching of the templates number with that of the word. If the device is capable of having a vocabulary of sixty four words, as is the Votan VPC 200, then one could type a twenty word list in three times, varying the order each time.

Study Nine

Some devices offer the opportunity to have words presented automatically or at the users' pace. Study nine was carried out to investigate the difference between these two modes of presentation. Subjects were either allowed to pace the enrolment process themselves, by pressing a key before speaking a word, or were forced to work at the pace set by the device.

Two measures of performance were used. The first was the recognition score provided by the device. The second was subjects performance on a 'surprise recall' test. It was hypothesised that the subjects in the user paced group would attain better results on this test because they had to process the words at a deeper level than the device paced group. The fact that subjects in the user paced group were required to cue their own speech means that they needed to pay more attention to the list of words than the device paced group.

Ten undergraduate students (eight men and two women), from various disciplines acted as volunteer subjects. They were assigned to either the user paced or device paced enrolment groups. A vocabulary of 20 items, relating to vehicle inspection, was enrolled on a Votan VPC 2000 installed on an IBM PC II.

The user paced group had to press the <return> key, to produce an acoustic beep, before speaking each of the words. Once the word was accepted, they could press the key to produce a prompt for the next word. All words in the vocabulary remained on the screen during the enrolment. The device paced group used the "Autotrain" function of the Votan. This allowed words to be prompted as soon as the preceding word had been accepted.

On completion of the first enrolment, both groups were given a recall test, to ascertain how many of the enrolled word had been remembered. The procedure was then repeated twice more, and results analysed using a Mann Whitney U test.

Results

Recognition scores were very similar for both groups (89.45, on average), and were not. statistically, significantly different. Figure 13.1 shows the results of the recall tests. These results were not found to be significant, but show definite learning effects for both groups (as one would expect), and an advantage for the user paced group on the first tests.

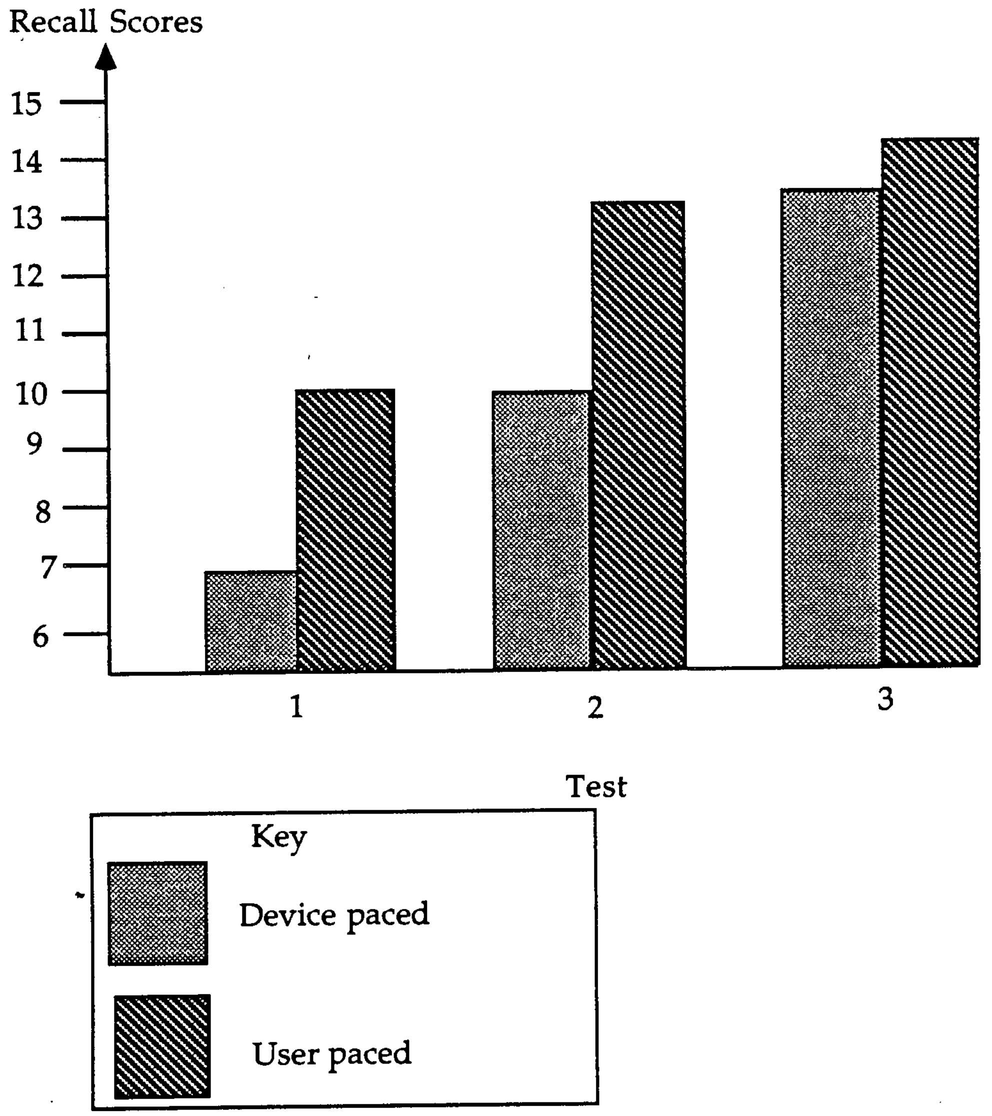


Figure 13.1: Results of recall test in Study Ten

Discussion

It was found that on a 20 word vocabulary, there was little difference in recognition accuracy over three passes. The results from the recall test suggest that the user paced group may be processing the words at a slightly deeper level than the device paced group, perhaps because they need to attend to their position in the list, whereas the device paced group have words cued individually. As the discussion of varying word order above suggested, providing users with a task requiring more attention than merely repeating words can help in their performance.

Enrolment and styles of speech

Teja and Gonnella (1983) advise users to enrol and use the ASR device in a monotonic voice. From the discussion so far, it is clear that this proposal is ill conceived. Not only is a monotone unnatural and difficult to use, it bears no relation to how users actually speak to an ASR device (cf. chapter ten).

Obviously before enrolment begins, users must be trained as to how best to use the device. The section on training below, shows that if users have an adequate model of how to speak, they can produce good representative templates during enrolment. Such templates can capture the style used to issue commands to the device.

Matching enrolment to working conditions

McCauley (1984) emphasises that enrolment should occur in an environment as similar to the working one as possible. People tend to speak differently in noisy conditions than quiet (presumably for the purpose of monitoring their speech). If this is the case, then direct feedback from the microphone via headphones will reduce the perceived need to shout. Otherwise, enrolment should be carried out in the same level of ambient noise as the proposed working conditions. Kersteen and Damos (1983) compared use of ASR in noisy conditions when enrolment had taken place under low or high ambient noise. Enrolment under low noise gave 78% performance, whereas that under high noise gave 97% when the device was used in high noise.

'Hidden enrolment'

Some writers suggest that 'task like' templates could be collected by disguising the enrolment process. If one is using connected word recognition, then words spoken will be effected by the words preceding or following them,i.e. the phonetic context. Enrolling devices by speaking individual words at a time will not be able to reflect this. Also, words which are being used to perform a task will be assigned task specific meanings,i.e. the semantic context.

Green et al (1983) propose that rather than prompting <say 'up'>, the user be prompted <mover the cursor up>. In both cases, the user will be required to say the word "up", but in the latter case the word is assumed to be spoken in a task like context. This scenario will work providing the user knows the vocabulary and is aware that he is dealing with an isolated word recogniser. It could be imagined that a user prompted with <move the cursor up> responding with "could you move the cursor upwards,please". Although exaggerated, this shows a potential problem. Such an approach is superficially attractive, but has limitations. Recognisers can be run either in enrolment or recognition modes, but rarely can both modes be run together. If a task is to be performed using ASR, then a word must be recognised. Words cannot be recognised before they are enrolled, nor can they be enrolled and recognised at the same time.

This means that for hidden enrolment to be used with an actual ASR based task, software needs to be written. Talbot (1986) demonstrated that task like enrolment could be carried out by asking subjects questions, which they answered according to a script, e.g. to 'where do you wish to travel to?', subjects were required to reply with the name of a town. Their utterance was used to create a template of that word. Such a procedure is only applicable for selected applications such as database enquiry.

It does not seem that the cost of writing such software can be justified by potential benefits, if it can be written at all. The use of simulations and games could help to make enrolment less arduous and provide useful familiarisation, but, as this study shows, enrolment can be made effective by using a demonstration to provide both a semantic and a functional context for the vocabulary.

Template adaptation

Meisel (1986) defines two types of enrolment which offer the potential of reducing enrolment time and allowing device adaptation. The first is 'extrapolative enrolment'. This involves defining subsets in the vocabulary which contain words with the same root, and extrapolating from the enrolment of a root word to the whole subset. For instance, if the vocabulary contains the words 'copy','copier', 'copies', then templates could be collected for the word 'copy' and extrapolations made to the other words. This seems a reasonable suggestion and is feasible for many devices.

However, when designing a vocabulary for ASR one attempts to keep the words as distinct as possible. 'Extrapolative enrolment' would produce a high potential for confusion between words by having several words with the same root. Also, it is difficult to think of that many words which share the same root in any particular application. In general, different actions require different commands to carry them out, and commands are usually issued in the present tense.

A more promising form of enrolment described by Meisel (1986) is termed vocabulary independent enrolment. Here the text used is independent of the application, and enrolment is based on the collection of phonemes from these words. The phonetic engine developed by Meisel and his colleagues uses a set of rules to combine these phonemes into legal; vocabulary items. Such an approach is exciting, but will require extensive research before it becomes commercially available. Chapter four discusses the issues surrounding phonetic based speech recognition.

One method of compensating for the inherent variation of speech is to provide some means of allowing templates to adapt over time to changes in the speakers voice. Thus, speech input is not only compared with the templates, but is used to provide a reference for updating the templates. McInnes and Jack (1987) describe an adaptation algorithm which was initiated by the user. Each time a word was misrecognised, the user had to acknowledge whether it was the word she had spoken. If the recognition was correct, a weighted averaging procedure based on dynamic time warping, dragged the template towards the analysis of the input word; if the recognition was incorrect, then the template was dragged away from the input word. Disappointingly, initial results showed no difference between adapted and nonadapted templates. However, a problem one faces with template adaptation is that any alteration to the template might make it similar to other templates in the set. In other words, an adapted template might actually be similar enough to another template to produce confusion. Therefore, some form of compensation was introduced by McInnes and Jack (1987). This took the form of providing an indication of the number of times a template had been adapted. Adaptation with compensation provided an increase of 2.7-2.9% in recognition accuracy.

It is important to note that the adaptation had to be supervised by the user. When unsupervised adaptation was used, the recognition accuracy decreased markedly. This shows that the linguistic knowledge of a typical user of English is essential for adequate adaptation algorithms. Research by Green et al (1983), and Damper and MacDonald (1984) reached similar conclusions. Template adaptation could improve the performance of ASR devices, but they are incapable of adapting templates without supervision.

Damper and MacDonald (1984) emphasise the need for the stability of the updating process. At the present time, the only means of providing stable feedback concerning the accuracy of speech recognition is through the user. In other words, before ASR devices can be self adaptive, they need to approach the capacity for language use of humans. This leads to a somewhat paradoxical argument; in order to improve performance, ASR devices could update templates using an adaptation algorithm; in order to to use an adaptation algorithm, without human supervision, ASR devices need to improve their performance.

The problem of template drift has already been noted. Baker and Pinto (1986) suggest the use of four templates per vocabulary item. Updating would take place on one template. The templates would be averaged to provide a score against which to check the updated template. If the deviation was too great, the template would be brought back towards the average value. Again, adaptation would be supervised by the user.

Ultimately, template adaptation will lead to speaker independent devices. The knowledge about what constitutes appropriate 'guesses' as to the words said, will enable the device to deal with any speaker. This appears to remover the need for enrolment. Hapeshi and Jones (1988) point out that,

> "true speaker independence of intermediate or large vocabularies is probably not a practical proposition since a degree of template training or adaptation will always be necessary."

So called speaker independent devices still require reference templates to be installed. These are recorded from a large number of speakers and averaged to give samples. The enrolment procedures outlined here will be useful to existing ASR devices. However, it appears more desirable to eliminate or at least reduce the burden of enrolment. Speaker independent devices are currently available. These devices do not require enrolment by the end user as they have a set of templates tuned to standard utterances of the vocabulary items. The standard utterances are based on average templates from several hundred speakers.

The disadvantages of speaker independent devices currently outweigh the advantage of not having to enrol them. The available vocabularies are usually small and predefined by the manufacturer. This means that the vocabularies are often standard sets of words, e.g. numbers and commands for telephone dialling. It would be better if they allowed some application specific tailoring.

Re enrolment

A viable alternative to both speaker independent devices and template adaptation, uses reiterations of the enrolment process. If a word is consistently misrecognised, this presumably results from a large mismatch between template and speech. Re enrolling the word ought to allow a better match. This would result from a more representative sample of current user speech being supplied, and could also replace a template corrupted by noise at enrolment.

The idea of re enrolling to provide representative samples of current speech has prompted some researchers to suggest that users re enrol the device on a regular basis. Danis (1989a) found that if users employed templates enrolled two weeks prior to the test, their error rate was 10.3%, compared with 6.9% using templates recorded immediately prior to use (a difference of 3.4%). Kohda and Saito (1973) found that error rate increased by 0.8%, when twenty days elapsed between enrolment and use. This illustrates a basic problem for direct comparison of performance using ASR devices; different devices use recognition algorithms of varying robustness. However, it does show that human speech variation over time will effect ASR performance.

Variations in human speech are very common. People do not speak the same word in precisely the fashion on any two occasions. However, current ASR techniques are sufficiently robust to capture enough of the constant properties of the speech signal to allow recognition to take place (see chapter four). As the studies reported above show, speech variation can still effect recognition performance over lengthy periods of time. While some forms of variation appear to be inherent in the speech signal (chapter four), other forms may be due to poor use. Speakers may be careless in the manner in which they enunciate words, they may speak too quickly for the device, they may attempt to 'help' the device by speaking with unnatural stress and emphasis. Such problems could be overcome by ensuring that prospective users received adequate training, which included an idea of the capabilities of the device, and by updating templates on a regular basis (Danis, 1989b).

A further form of variation in ASR performance has been noted. Rather than being caused by the variations of speech over time or by careless use, some researchers have noted that performance can degrade over a single session of using ASR. Zarembo (1986) found that performance decreased dramatically over a days use. He attributed this finding to the vocal fatigue of the users. However, telesales staff are required to speak for up to eight hours a day. Their work requires short telephone conversations to be repeated several hundred times a day, and incidents of reported vocal fatigue are very low (Yellowpages, 1990). This suggests that the findings of Zarembo (1986) may be due to another phenomenon.

Frankish et al (1990) have shown that error rate rises from an average of 7.9% to an average of 10.7%, over a task lasting half an hour. This result was replicated on later trials, which suggested that a factor beyond learning the task was involved. By asking subjects to pause briefly for short periods during the task, Frankish et al (1990) hoped to counter any effects of fatigue or boredom. However, they found that the error rate still increased.

In their third study, Frankish et al (1990) found that, if subjects re enrolled the vocabulary half way through the task, their error rate dropped significantly. This was used to argue for the presence of a 'warm up' period, in which users settle into a stable speech pattern. If this is the case, then some period of 'warm up' needs to be incorporated into all ASR systems. However, the task used by Frankish et al (1990), required users to speak place names at computer generated prompts. Enrolment took the form of reading the list of names, prior to use. As was discussed above, the differences between enrolment and task styles of speech are important factors in ASR training. The fact that different styles of speech were used in the two aspects of the studies may be a contributory factor.

Frankish et al (1990) rightly point out that their results stem, not from vocal drift (a drift from template and present speech), so much as vocal instability. Subjects took a period a time to restabilise their speech style. This suggests that some learning was, in fact, involved in the task; with users having to adapt to the recognition requirements of the device. If this was the case, then why did the evidence of a 'warm up' effect persist through several trials? If subjects had learned a correct speech style, why were they unable to remember it in future use? The answer to these questions lies, I feel, in a consideration of the type of knowledge users employ to represent the task. This is considered in the section on user training below.

User Training

Mc.Cauley (1984) has noted that user training is,

> " largely underestimated in the development of speech interactive systems."

Anyone who has used an ASR device will know that, contrary to manufacturers' claims, it can seldom be used perfectly first time. Rather, some hours of practice are required in order to develop an appropriate style of speaking. The variability of human speech is widely recognised as one of the major problems for ASR (Lea, 1980; Martin,1976). This variability manifests itself in the inability of some users to pronounce the same word in the same fashion on different occasions (Doddington and Schalk, 1981). Such problems of speech consistency are increased when users try to use connected word recognisers.

ASR devices can recognise either isolated words or short, connected phrases. Neither of these styles of speech is used to any great extent in normal conversation. If an attempt is made to speak to the device using conversational speech, its performance will decrease dramatically. Thus, users of ASR devices do not employ a 'natural means of communication', as promoters of ASR often claim, but must learn to adapt and develop their speech to fit the requirements of the device. Rather than starting to speak to a computer, users must adapt their speech to the devices particular constraints. Humans are able to adopt different styles of speech according to their perceptions of different social situations. Speech based interaction with machines can be regarded as a novel social situation which requires a particular speech style.

Not all users are able to adapt to the use of ASR efficiently. Around 90% of new users are able, with practice, to achieve optimum performance (Doddington, 1980). The remaining 10% have problems. Doddington and Schalk (1980) draw a distinction between users who are capable of using ASR (termed 'sheep'), and those who are not (termed 'goats'). The presence of 'sheep' and 'goats' can be present even after training and four weeks of practice (Danis, 1989a). This means that it is a serious problem.

There has been little research into how to resolve the problem of 'sheep' and 'goats'. That is, how to assist the 'goats' to achieve optimum performance. From observation of initial users in undergraduate classes, I suggest that two views guide initial practices. Some users approach an ASR device with the assumption that it will perform well, that it will do exactly what they tell it, that they do not need to alter their normal speech to accommodate it, and that errors are due to the device 'malfunctioning'. In general these are the users who do not have problems, beyond normal error rates of around 2%.

Other users are more tentative. They try to speak to help the device and assume all errors are due to poor speaking. These users perform very poorly. The distribution in the classes reflect the 90% 'sheep' and 10% 'goats' split distinguished by Doddington and Schalk (1981). The performance of students in the class suggests that differences in performance may be due to attitudes to the device. Furthermore, these attitudes seem to reflect beliefs in who is in charge of the interaction, such as whether it is the user who should accommodate the device or vice versa. This could be assessed using the notion of 'locus of control' (Rotter, 1966), to indicate whether the 'sheep' and 'goats' distinction reflected such beliefs.

One could hypothesise that the results from a study employing 'locus of control' to test users attitudes to ASR , will only capture some of the factors which make for poor users. If prospective users receive sufficient training, it is difficult to conceive that they will not be able to perform adequately. In Study ten , it is proposed that ASR use is a skill and consequently, needs to be taught appropriately.

The importance of training in learning to use ASR

If it is accepted that using ASR is an example of skilled behaviour, then training and practice are very important in its development. Personal experience suggests that minimal training can be supplemented by several hours of practice to achieve reasonable performance. But such lengthy practice can be expensive and time consuming. Estimates of practice time needed to reach acceptable performance levels range from three hours (Poock, 1980) to two weeks (Cochran et al 1980). Danis (1989a) found that error rate dropped by 33% over a four week period. In her study, subjects not only practised with the equipment, but also received instructions on alter their "speech habits" to improve performance.

The benefits of lengthy practice will obviously be affected by the users' level of motivation to use the device. The necessity for practice can be reduced by providing new users with some form of training. Many writers, e.g. Cochran et al. (1980), suggest that new users be prepared by briefly explaining the working of the device in an effort to appreciate its limitations.

Wilpon and Roberts (1986) have shown that subjects who are given instructions "regarding successful interaction" with an ASR device performed better than subjects who received no instructions. In a second study, additional instructions on correcting recognition errors were shown to improve performance even more. This suggests that device specific instructions are effective and helpful to new users; the user cannot simply speak to the device and expect it to perform well, despite the apparent naturalness of speech as a means.of communication. However, Wilpon and Roberts (1986) used a simple recognition test on an isolated word recogniser. If ASR is to be used effectively in industry, then studies must be carried out in which ASR is used to perform specific tasks. With such studies come the problems of defining adequate measures of performance.

The use of ASR as a skill

One of the reasons why a user cannot simply use ASR on meeting it, is that speech is an 'overleamt' skill. Such skills have become automatic as the result of practice (Anderson, 1980). Experienced users of ASR are often surprised that new users have difficulty. It is easy to forget that speaking to machines is not 'natural', but requires a particular style of speech, i.e. short, succinct, task related phrases with determiners and pronouns removed (see chapter ten). Thus, a particular skill needs to be adapted to a new set of requirements. This process of adaptation could be carried out in two ways. Either the users are given an ASR devices and told to practice until they reach optimum levels of performance, or they are trained.

It is suggested that speech is an 'overleamt' skill. Giving people instructions to speak in a specific style will require them to introspect on this skill. There are two reasons why it is difficult for people to do this. The first reason is common to all types of skilled behaviour. One of the characteristics of skilled behaviour is that it is automatic, i.e. one is not consciously aware of the various components which make up the skill (Fitts, 1964; Annett, 1989). The second reason why it is difficult to introspect on the process of speaking is that it requires the workings of low level cognitive and physiological systems which are unavailable to conscious introspection (Johnson-Laird, 1983).

Introspecting upon skilled behaviour can disrupt performance. For instance, asking someone to describe a shot as they play it in squash can disrupt their performance of that shot. Moreover, they might find that many of the actions cannot be put into words. This suggests a distinction between two types of knowledge. Such a distinction has been demonstrated in neuropsychology. Cohen and Squire (1980) have found that while amnesiacs cannotremember verbal facts, they could retain motor skills.

Fitts (1964) proposed that observational learning played as important a part in learning new skills as verbal instruction. Learning to play a shot is easier if one can watch someone else play it, than to receive instructions. Not only does the demonstration convey more information than written or spoken instructions, it also puts the actions in a suitable context for them to be learned, i.e. playing a shot in squash. Thus, it is possible to distinguish between information which can be put into words, and that which needs to be demonstrated. This echoes the distinction between declarative knowledge, i.e. facts that we know, and procedural knowledge, i.e. the skills that we know how to perform.

Providing instructions on how to use an ASR device will provide some declarative knowledge, e.g. how the device works, what task the user is required to perform, and a description of the procedure to be learned, i.e. "speak in a clear, consistent voice as if you were speaking over a noisy telephone channel". However, the provision of such knowledge does not necessarily entail the rules for its application. As there is little possibility of introspecting on the processes involved in articulating speech, it is difficult to propose the rules for speaking clearly and consistently. Advice such as 'don't mumble' could be offered, but this will only sketch the desired rules by suggesting a tentative set of boundaries. Therefore, it is necessary to determine either the type of conscious knowledge used in speech, or means of training behaviour which cannotbe put into words.

Types of knowledge required for ASR Use

A number of studies have shown that providing subjects with knowledge of the process they are required to control, does not effect their performance (Shepherd et al 1977). Mann and Hammer (1986) found that providing subjects with a set of procedures combined with theoretical principles of the process operation produced more errorful performance than procedures alone.

Morris and Rouse (1985) found that subjects provided with procedures showed more stable control of a process than subjects provided with principles of process operation. These studies suggest that procedures give operators an effective strategy, which cannot be communicated with the use of theoretical principles.

Broadbent (1977), Broadbent et al (1986) and Berry and Broadbent (1988) show that subjects who were allowed to practice on simplified decision making tasks improved in their ability to make the right decisions, but not in their ability to answer related questions. Berry and Broadbent (1984) show that verbal instruction can improve subjects' ability to answer task related questions without effecting their performance on the task.

These results can be related to the distinction already drawn between declarative and procedural knowledge. Some knowledge can be put into words, whilst other types of knowledge cannot. It is possible to draw a parallel between declarative knowledge, and the knowledge which can be put into words in these studies (termed explicit knowledge by Broadbent et al 1986). It is not possible to answer questions without being able to put the requisite facts into words. However it is not so easy to draw a straight forward association between procedural knowledge and the knowledge which is used to perform tasks (termed implicit knowledge by Broadbent et al 1986). There are some aspects of procedural knowledge which can be put into words, but implicit knowledge is difficult to articulate and is largely intuitive. It is suggested that the distinction between declarative and procedural knowledge be expanded to contain the notion of implicit knowledge. This can be thought of as a continuum across the range of knowledge and skills and the extent to which they can be verbalised.

From the studies reviewed, it appears that verbal instruction or training based on theoretical principles may not be the most effective means of training, in some cases. As Broadbent et al (1986) point out,

> "From the present state of knowledge, therefore, it would be unwise to assume that verbal knowledge is the ideal towards which the less explicit intuitive decision is developing. Rather, performance based on verbalisable knowledge and that which selects action by matching the situation to those met in earlier experience, may be alternative modes of function each with its own disadvantages."

Different tasks can be shown to have different levels of implicit and explicit knowledge. If using ASR contains a significant level of implicit knowledge, then providing instructions will not effect subjects' performance. A demonstration could be said to provide 'practice by proxy', allowing subjects to receive the relevant implicit, procedural knowledge.

Instructions and demonstration as media for training users of ASR

Telling a person to speak consistently does not provide salient explicit information, it only makes them aware of the possibilities contained in the word 'consistently'. This would suggest that if the task of using ASR be presented in demonstration the number of possibilities the subject is faced with will be reduced. Rather than considering what is required in order to speak consistently, the subject will be able to copy a particular style of speech.

By giving subjects a demonstration, it is possible to bypass the problem of deciding what constitutes explicit or declarative knowledge for using ASR. Subjects are provided with the requisite procedural knowledge. The use of ASR will be placed in the context of performing a task and should be easier to learn.

These hypotheses relate to Fitts' (1964) three stage model of learning cognitive skills. In the first, or cognitive, stage a description of the procedure is learned. In the second, or associative, stage a method for performing the skill is worked out. The second stage eventually merges into the third stage in which the skill becomes automatic (c.f. Anderson, 1980).

From this definition, providing a demonstration of the procedure to be learned offers a higher entry on the ladder towards automaticity. Rather than having to consider and attempt to apply facts concerning the correct use of ASR, subjects will be able to mimic an observed set of actions. It is expected that subjects who receive a demonstration will achieve better performance than those receiving instructions. Study ten investigates these issues.

Study Ten

Data were obtained from 11 students at Aston University: 6 in the demonstration Group (Group D); 5 in the instructions Group (Group I). The study used a speech based telecommand demonstration described in chapter nine.

As the vocabulary was known to contain possible confusions, we designed two methods of error correction which were simultaneously active throughout the task. Subjects could either repeat a misrecognised word or the whole command phrase. Further, a 'retrain' facility allowed subjects to escape from the task and update templates of problematic words. The desired word can then be selected and retrained. The user can then return to the telecommand page to resume issuing commands.

In the demonstration condition (Group D), the experimenter opened and closed five switches by issuing spoken commands using the telecommand 'syntax'. Any misrecognised commands were corrected, and one command word was updated. No instructions were given to the subjects during the demonstration. In the instruction condition (Group I), the experimenter explained how the device worked. The subjects were shown some pictures of previously recorded templates, and told how they were made. Care was taken to emphasise how important it was to speak consistently to ensure adequate matching of templates and speech.

All subjects were guided through the enrolment process and receive an explanation of the task. Subjects then perform five practice openings or closings of switches, before carrying out the experimental task. This was to change the status of fourteen switches.

When subjects had completed this task they were given two assessment sheets to complete. The first sheet contained a series of statements with a choice of two words for the subjects to select, e.g. "The system was EASY/ HARD to use". The second sheet contained a 15 statements. Subjects were asked to rate their level of agreement with these statements on a Lickert scale.

It was decided to use a performance measure related to the actual task. In order to issue a telecommand, subjects need to speak five words, i.e. "At Moseley 400kV. close P808". If the device recognises all these words first time, then the command will be issued in five words. However, if a word is misrecognised, the subject repeats it to complete the command, and the number of words used increases from five to six. Acceptable performance limits were defined as subjects not having to repeat more than two words per command (see study three, chapter nine).

Results

The mean distance score was very similar for both groups: 55.18 for Group D, and 54.8 for Group I. But this does not reflect the marked difference in performance between the two groups. Group D did not need to retrain any words while carrying out the task. Group I had to retrain around 38% of the words, on average. Consequently, Group I took longer than Group D to complete the task. (mean time for Group I =814.7s; mean time for the Group D = 304.6s

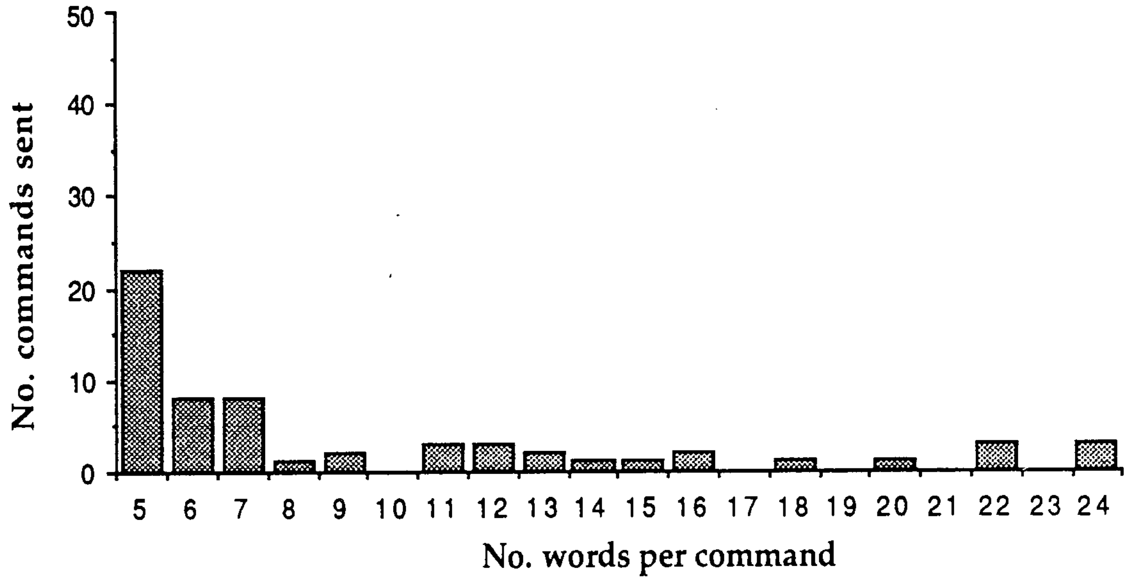


Figure 13.2: Number of Words per Command for Group I

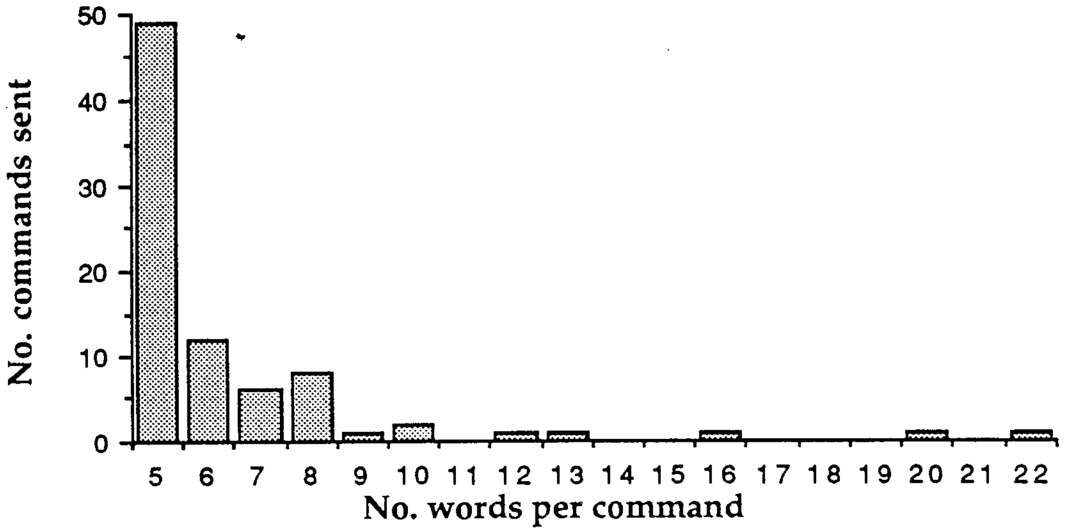


Figure 13.3: Number of Words per Command for Group D

The fact that the Group D had to retrain words meant that they completed less telecommands in acceptable limits. Using this criteria, the Group D issued 81.7% of its telecommands within acceptable limits, while the Group I managed 56.7%. Figures 13.2 and 13.3 show the distribution of words per command over the two groups. The difference between the groups in terms of the number of words used per command was analysed using a chi-square test, and was found to be significant. ($p<0.005$).

From their responses to the assessment sheets, one can conclude that Group D unanimously found the system easy to use, and rated the recogniser as performing better than they expected. There was some disagreement as to whether the device had made any errors, with 3 subjects claiming that the device had made none. All subjects in this Group found the number of errors small and acceptable.Results from the Group I were less clear cut. Four subjects rated the device as performing worse than they expected, and three subjects found the device hard to use. All subjects said that the device made some errors. It is noteworthy that all subjects did not feel that they spoke in the same way all the time.

Discussion

For the purposes of industrial based ASR systems, users tend to use speech purely to issue commands, as opposed to conversationally (see chapter ten). Group D were able to observe speech being used in this manner. Group I received instructions on how to speak in this manner. The performance results suggest that the main difference between the two groups lay in the adequacy of their templates.

When they enrolled, subjects in Group I were conscious of trying to create 'good' templates (as several subjects mentioned during debriefing). The creation of 'good' templates could result from a number of factors. Subjects were attempting to speak 'normally', but were not sure how their normal speech sounded. Instructing them to speak 'consistently' or 'normally' encouraged them to concentrate on the sound of their speech, rather than on the task of issuing telecommands. Group I knew that the words were used to control a telecommand demonstration. They knew what the words meant and what they were used for. In other words, they were provided with a semantic context for the vocabulary.

This would constitute a set of explicit facts concerning the vocabulary, i.e. declarative knowledge.

One of the subjects in Group D said that she was aware that the words she was enrolling would be used to perform a task. This suggests that she was able to form an idea of what constituted an appropriate way of speaking the words, such that her utterance of them during enrolment would be similar to that during the performance of the task. This appropriate style of speech results from the subjects having knowledge of the context and meaning of the words. Group D knew what the words meant and how to use them. However, in addition to having a semantic context, subjects in Group D also had a functional context for the vocabulary, i.e words were associated with specific actions.

Thus, the demonstration provided them not only with declarative knowledge concerning the device and the task, but also a set of rules for applying this knowledge. It is not certain how much of the declarative knowledge subjects received from the demonstration because they were not questioned after the experiment. Their performance of the task, and the consistently high performance scores suggest that they were able to assimilate sufficient implicit, procedural knowledge to be able to speak consistently to the device. This led to a more consistent match between recorded templates and task speech, which in turn led to better performance.

Study ten suggests that in order to use ASR efficiently, users should be provided not only with a semantic context for the vocabulary, but also a functional one. Whilst some form of task like enrolment can provide this expanded context, this study shows that demonstration by an experienced user of ASR is effective in providing such an expanded context, yields better performance than the use of spoken instructions, and is easier to provide than such task like regimes.

This finding is explained by the notion that a demonstration provides subjects with the requisite implicit, procedural knowledge which may not be verbalisable. Attempting to put such knowledge into to words may even be disruptive to performance, because it forces subjects to concentrate on inappropriate aspects of the task. Therefore, it is recommended that where ASR systems are installed, users are provided with a demonstration as part of the training programme.

CHAPTER FOURTEEN

THE USE OF ASR IN SITUATIONS OF HIGH COGNITIVE WORKLOAD

This chapter investigates the use of ASR in high workload environments. It presents a discussion of the notion of compatibility, used in control room design, and argues that the use of ASR requires consideration of cognitive compatibility. Cognitive activity is discussed within the framework of human information processing. A brief discussion of theories of human attention leads to the conclusion that people use at least two codes for information processing: verbal and spatial. Cognitive compatibility will require that a task does not involve the translation between codes, in order to maintain operator efficiency. A review of research investigating the issues surrounding this notion is presented, with particular reference to the work of Wickens and his colleagues. Cognitive activity can be studied using dual task experiments. Study eleven investigates the affect of spatial and verbal secondary tasks on a primary task requiring speech based control of a simulated chemical process. The results indicate that ASR use is a verbal activity which can be paired with spatial tasks without detriment. This supports the proposal that ASR can be used in control rooms for command and control.

Introduction

Chapters six to nine presented current and potential uses of ASR. Decisions concerning when best to use speech recognition are judged primarily in terms of physical task demands. For example, baggage handling can be performed more easily if the operator performs the requisite data entry using ASR than if he used a keyboard (see chapter six). This example illustrates the potential of ASR to allow eyes free interaction with a computer, when the operators hands are busy and when the operator is moving around the computer unit. However, as we saw in study one, allowing 'eyes free' operation with ASR may actually hinder operator performance. Thus, it is not enough to consider the use of ASR simply in terms of physical activity.

ASR could present advantages, over manual data entry, in cognitive tasks. For example, in the baggage handling task, the operator needs to retain the number or destination on the bag in short term memory, between reading it and typing it into the host computer. It is well attested that short term memory has a limited capacity, and that data in short term memory is subject to errors in recall. As the complexity of the task increases, the possibility of errors in recall will increase. The capability for direct data capture offered by ASR reduces this simple level of cognitive workload.

Welch (1977) proposes ASR will result in better performance of 'complex' tasks, than manual data entry. In chapter seven, this claim was redefined to argue that ASR would be useful in complex, verbal tasks, such as command entry. Study three (chapter nine) supported this claim by employing ASR in telecommand. Therefore, ASR can offer benefits in terms of cognitive activity. However, the nature of the cognitive demands which ASR places on the user in control room systems has not been investigated. Study eleven will address this issue. From this we will be in a position to propose how ASR use will interfere with the activities of the control room operator identified in the Introduction.

Compatibility

Study seven showed that verbal decisions are best aided by textual feedback to the operator. Study eight showed that overall plant control was best aided by a combination of textual and symbolic feedback. These results suggest that certain tasks are best supported by specific types of feedback. In other words, a degree of compatibility exists between the information processing requirements of the operator and the feedback provided, during the performance of different tasks. This conception is by no means novel (human factors research has long shown that operator performance can be optimised by pairing 'natural' responses to specific stimuli or task demands), but it remains to be sufficiently investigated in control room design.

There are several types of compatibility. The most obvious is movement compatibility. Imagine a knob which can be rotated to move a cursor on a screen. One would expect that rotating the knob clockwise would produce a corresponding move to the right by the cursor. There is no inherent reason, in terms of the systems mechanics, why such a relationship should hold, but it seems natural to the user. With practice, it would be possible to operate a cursor by moving the knob in the opposite direction to the 'natural' one. However, there are a number of problems with this, not the least of which being that under stress, people tend to revert to natural actions. For example, a well known story in human factors concerns an operator of a large press. In order to lift the press, one pushed a lever down and in order to lower the press, one lifted the lever up. Such a design seems simple enough, and conforms to basic physical principles of leverage. However, the operator spotted a piece of metal becoming lodged in the press and decided to raise the press as quickly as possible. He lifted the lever with all his might, and brought the press crashing down. This resulted in a wrecked press, and is a sobering lesson in the consequences of violating 'natural' movement compatibility.

A second type of compatibility can be defined as spatial compatibility. This is best illustrated in a classic study by Crossman (1956). Subjects were faced with a panel of lights and had to press corresponding buttons as soon as the lights appeared. The buttons were arranged in a row, numbered one to eight. The lights were either in a random order (symbolic condition) or placed directly above the corresponding button (nonsymbolic condition).

Crossman (1956) found that the condition in which the lights appeared above the corresponding button produced much faster reaction times than the random order condition. This was assumed to result from the fact that the subjects in the random order condition were required to translate the lights number to the appropriate button number, whereas the other condition could use the more obvious relationship of light and button.

This deceptively simple study highlights a basic principle of human factors: while the time to make decisions increases proportionally to the amount of information presented (Hicks, 1952), high compatibility mappings can bypass, or at least reduce the necessity for, information processing (Fitts and Posner, 1967; Kornblum et al 1990). High compatibility could save valuable time in emergencies and could reduce the likelihood of human error. One could ask whether such compatibility will exist for cognitive tasks; after all, the symbolic condition used by Crossman (1956) appeared to require some cognitive operation by the subjects in order to translate the lights number to the correct switch.

Brainard et al (1962) used two types of stimulus and two types of response to determine whether a relationship existed between the way manner in which information is presented and a particular response. In one condition, a set of numbers were presented auditorially and subjects were asked to either speak the numbers or press corresponding number keys. In the other condition, numbered lights were illuminated in defined sequences. Subjects were again required to speak the number or press an appropriate key.

Results showed that performance was best when auditorially presented numbers were paired with a spoken response. Next came lights with key press. Thus, there is a level of compatibility between speaking and auditory presentation, and manual response and visual presentation. Greenwald (1979) defines this as 'ideomotor compatibility'. He shows that a spoken response to a heard letter or a written response to a seen letter, produces faster reaction times than conditions without ideomotor compatibility (such as a written response to a heard letter, or a spoken response to a seen letter).

The results of these studies suggest that information processing demands can be reduced by using some form of compatibility (movement, spatial, ideomotor). This concept was developed in detail by Fitts (1954), who proposed that optimal performance will result when stimulus characteristics are compatible with response demands.

The principles behind stimulus response compatibility become hazy when one asks why the pairing aids performance. Fitts (1954) suggested that people use population stereotypes based on previous experience. This notion is difficult to define, although see Kornblum et al (1990) who present a detailed account of stimulus response compatibility research, and relates to the development of the concept of relationship between an action and a stimulus. Hick (1952) suggested that a series of subdecisions are made concerning the correctness of a response. The number and complexity of the subdecisions are related to the compatibility between stimulus and response. While this appears to be a parsimonious explanation for behaviour on a physical tracking task (Harvey, 1988), it cannot adequately address the complexity of cognitive activity. The concept of refining subdecisions in relation to a goal can be explained, in physical terms, as the discrepancy between goal point and state point. How does one begin to define this discrepancy in cognitive terms?

Stimulus Response Compatibility in the Control Room

The concept of stimulus response compatibility is applicable to systems which require manual control. When we consider process control operation, the concept needs to be extended. With the technological developments which are taking place in control room systems, the task of the operator is increasingly becoming one of monitoring and supervision, rather than direct manual control (see chapter one). This inevitably increases the cognitive demands made on the operator, and introduces the problem of how efficiently the operator can monitor system changes or make decisions concerning appropriate action. This suggests that,

> "The relevant model of the process controller...
> [is] a multichannel decision maker embedded
> within a fairly slow response dynamic system."
>
> [Edwards, 1976]

Due to the complexity of the process under control, operators need to gather information from several different sources to assist in their decision making, and to generate a number of different types of command to control the process. In the Introduction, the currently popular notion of 'mental models' was discussed. It was proposed that operators assimilate incoming information into internal representations of process state and activity. However, 'mental models' are incomplete and different 'models' are developed for different aspects of the process. This demonstrates that operators actions can seldom be regarded as simple pairing of stimulus and response. Rather data is manipulated and processed by the operator, in terms of his understanding of the current process state, to produce a,

> "tactical and strategic interaction with the
> environment in the search for optimal
> adaptive behaviour."
>
> [Moray, 1976]

This means that, in order to be viable for control room design, the concept of stimulus response compatibility needs to be extended to include cognitive activity. Such activity will necessarily mediate stimulus processing and response generation. Stimulus processing can best be defined as a form of attention, which is a topic central to cognitive psychology.

Theories of Attention

Operators in the control room are faced with a wealth of information. Obviously it is not possible to attend to all of this information, and so they must be capable of limiting the information they deal with. Although it is common practice to present the term "attention" as if it is unambiguous, it is not an east term to define. Current thinking in cognitive psychology questions its utility (Allport, 1980), seeing as being "too amorphous to be of much value." [Eysenck, 1984].

The following discussion rests on the assumption proposed by Kinchla (1980), that,

> "Rather than treating attention as a single entity,
> it seems more useful to assume that a variety of
> cognitive mechanisms mediate selectivity in
> human information processing."

In order to effectively describe the cognitive component of the stimulus response compatibility, it is necessary to define the probable form of these "cognitive mechanisms." A central argument in attention research concerns the selection of information with respect to processing. During the 1950's it was assumed that information was first selected by a form of filter, and then processed (Broadbent, 1958). Although simple and parsimonious, the filter theory was felt to be too limited. It was superseded by a theory proposed by Triesman (1960; 1964), which replaced the filter with a short term store which gradually attenuated the strength of incoming signals. The attenuation theory was further modified by Deutsch and Deutsch (1963), who argued that information was selected prior to processing on its relevance to the task in hand. Norman (1968) proposed that all incoming information had to be processed prior to selection. Information which was deemed irrelevant would not receive further processing, and be quickly forgotten. This would allow the explanation of subliminal pattern recognition.

The discussion has illustrated the complexity of attention. Some writers suggest that all the theories are applicable to psychology, depending on the task being studied (Barber and Legge, 1976). Alternatively, one could propose that the theories are applicable to different aspects of the task. This would require processing of some information to be carried out in terms of physical characteristics (Broadbent, 1958), some information to be processed in terms of the information context (Treisman, 1964), and some information to be processed in terms of the situation (Norman, 1968). The question of where to place the filter (which has exercised attention theory for several decades) is a red herring . Rather, one should ask how the filtering process can be varied. Johnstone and Wilson (1980) showed that subjects can vary their attention according to the demands of the task. This means that attention cannot use a fixed form of filtering. However, the concept of flexible attention cannot explain why it is impossible to attend to a large number of items. One must assume that attention can be distributed, within certain limits.

Kahneman (1973) proposed that there exists a general limit to a person's capacity to perform mental work. This capacity varies in accordance with the level of arousal required to perform a task, in accordance with the Yerkes-Dodson (1908) law. A major problem with the theories considered above is that they cannot explain how two tasks could performed concurrently. Kahneman's (1973) theory suggests that providing the combination of tasks does not produce too much arousal, they can be performed together. This theory is simple, and popular, but is flawed; it relies on a circular argument, viz. in order to measure limits of attentional capacity, one must ask subjects to perform two tasks concurrently. If the tasks interfere, it is because capacity is limited; the hypothesis explains itself.

There is a growing body of evidence which suggests that some activities can be performed automatically. That is, some skilled behaviours become sufficiently well practised to be performed with conscious control. These automatic processes cannot be held to represent a separate category of behaviour (Posner and Snyder, 1973), but all behaviour should be considered on a scale of automaticity (Shiffrin and Schneider, 1977). Automaticity cannot be accommodated by the attentional capacity theory. Finally, capacity theory assumes a gross level of processing. While such simplicity may be useful for general discussions, it cannot be used to predict what type of tasks will interfere (beyond the dimension of difficult and simple tasks).

A theory of attention based on the structure of the information processor seems more plausible than concepts based on filters or capacity limitations. In other words, one must distinguish between structures of processing which can be modulated in their availability with respect to task demand, and a central capacity of attention (Navon and Gopher, 1979). "Attention" can be seen to be more flexible than filter theories suggest. This flexibility leads to a degree of confusion in the use, and opacity of meaning, of the term. Allport (1980) has criticised researchers for using the term as a synonym for the equally vague term, "consciousness". Such use does not offer any explanation of how the process of attention works. As an alternative, Allport (1980) proposes a theory in which several distinct processing resources can be called upon simultaneously. Tasks which use separate resources can be performed concurrently, those which require the same resource will interfere. Eysenck (1984) has criticised this view, saying that. although it seems plausible, the notion of separate processing resources would lead to chaos and confusion without some form of central control.

This is not strictly true. Barnard (1985) describes a plausible theory of interacting cognitive subsystems, in which the subsystems are assumed to run independently. Each subsystem has a defined set of computational activities and processing constraints. They work together, 'self selecting' themselves at various points in the processing activity. This theory is still being developed, and draws in ideas from all aspects of cognitive psychology. It is very ambitious, but not currently workable.

While Allport (1980) and Barnard (1985) present models which seem plausible, it is difficult to use these models as theoretical tools. They are constructed post hoc, and cannot, as yet, be used to generate experimental predictions. However, the concept of several processing resources being available in a limited capacity attentional system does seem to be able to explain much of the experimental findings reported. What is needed is a more parsimonious version of such a theory.

This can be developed within the framework of working memory theory, developed by Baddeley and Hitch (1974). They begin by simplifying information processing into two systems: verbal and visual spatial. These systems are coordinated by a central executive, which can allocate attentional capacity between them.Verbal processing is carried out by an 'articulatory loop'. This in turn has two components, an articulatory control process which handles speech production, and a phonological store which maintains and stores verbal information (presented either visually or auditorially) for a duration of around two seconds (Baddeley, 1987). Visual spatial processing is carried out using a 'visuospatial scratch pad' which is a short term store for information relating to spatial location (Baddeley and Lieberman, 1980).

Multiple Resource Theory

We have seen how compatibility can exist across stimulus response pairings. It can also exist across stimulus processing and processing response pairings. This is the basis of the theory developed by Wickens (1980; 1984) and his colleagues: multiple resource theory. The basic principles of the theory can be represented pictorially, see figure 14.1.

Figure 14.1: The Structure of Processing Resources
[from Wickens, 1984]

Wickens model is designed to address cognitive performance. This means that it only employs two sense modalities: auditory and visual. Information is received in one or other of the modalities, and encoded as being spatial or verbal data. "Spatial" information relates to "the three axes of rotation or orientation" (Wickens et al 1983) and "verbal" information relates to "the use of language or some arbitrary symbolic coding" (Wickens et al 1983). After encoding, the information is processed using the appropriate processing code. This processing leads to the generation of a response, which is either spoken or manual. It is possible to develop a number of experimental hypotheses from this simple model, which relate to the general concept of compatibility. By maintaining a constant path through the model, that is, using components of the same processing code, it is assumed that performance will be better than if different codes are used. This relationship will hold for the performance of single tasks, and is illustrated by figure 14.2.

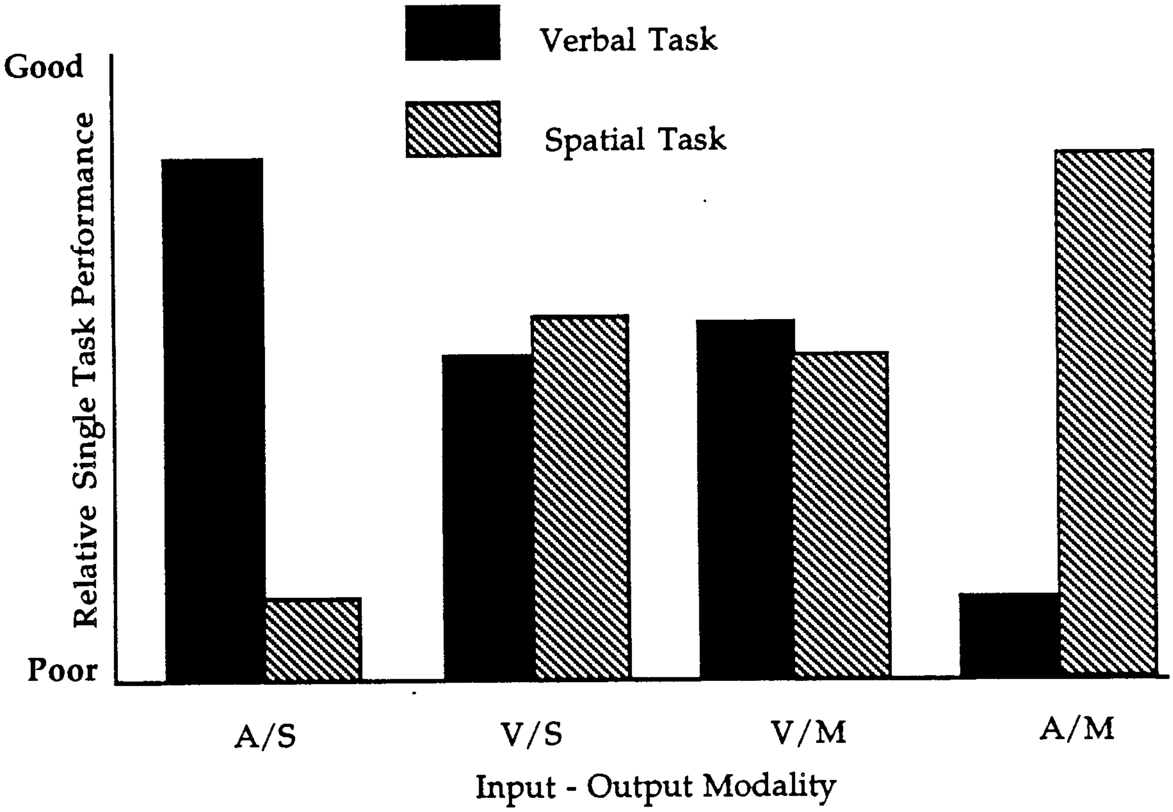


Figure 14.2: Single Task Performance Predictions
[adapted from Byblow, 1990]

For the performance of two or more tasks, the model shows where interference may occur, as shown in figure 14.3. The primary task is a tracking task, such as flying an aircraft, the secondary tasks are either verbal or spatial. The figure shows that highest interference occurs when a spatial secondary task, using manual response to visually presented stimuli, is paired with the primary tracking task.

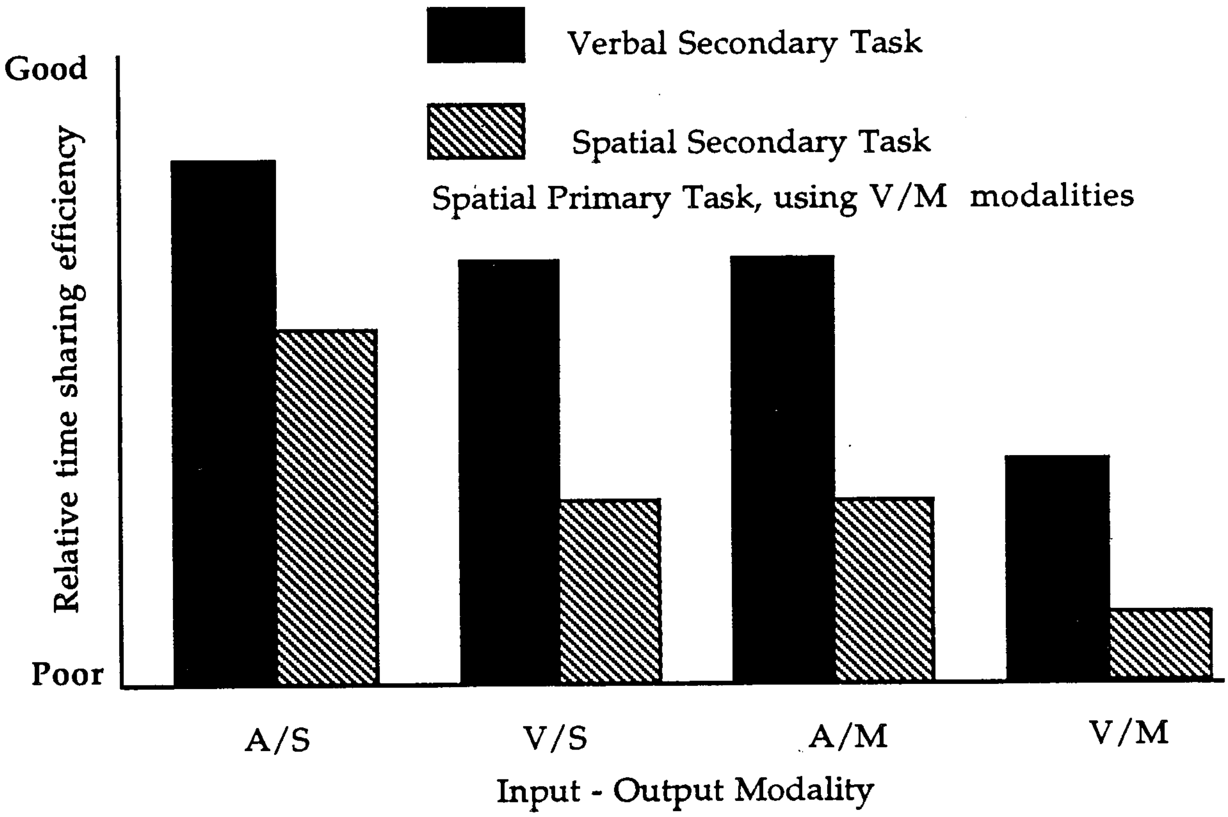


Figure 14.3: Dual Task Performance Predictions
[adapted from Byblow, 1990]

This shows that rather than generating simple classifications of activity, one must examine the total task structure in order to determine the potential areas of interference (Berman, 1986). However, in complex domains, such as the process control room, this is easier said than done. There are a number of variables which can combine to produce extremely complex and variable patterns of activity. This makes experimentation impossible. Consequently, research into multiple resource theory has largely been confined to the laboratory. Further, much of Wickens work has been concerned with the performance of fighter pilots. Tasks and cognitive demands in the fighter cockpit can only be tenuously related to the control room, and so it is necessary to begin investigating the theory in the context of control room systems. Before reporting a study of multiple resource theory in a control task, it will be of use to review several of the studies to highlight to benefits and problems of the theory.

Studies in Multiple Resource Theory

a.) Tracking paired with digit subtraction task
(Harris et al 1978; Wickens, 1980)

Tracking took the form of maintaining the position of a cursor in the centre of a CRT, using a joystick. The cursor moved off the goal point at random intervals and magnitudes.The digit subtraction task required subjects to subtract the second digit of a pair from the first. Digits were presented either on a CRT or over headphones, and subjects either spoke their response or typed it. Results showed that pairing the auditory display with the spoken response, on the digit subtraction task, gave less interference on the tracking task than other combinations.

b.) Tracking paired with memory search reaction time task
(Wickens et al 1983)

Tracking was the same as in (a.). The memory search reaction time task (adapted from Sternberg, 1975) required subjects to decide whether a three letter display matched a member of a previously learned set. Speed of response was measured. Displays were either via a CRT or headphones, and responses were either spoken or by 'yes'/'no' buttons. Results showed that, as in (a.), pairing the auditory presentation with the spoken response, on the memory task, gave less in interference on the tracking task than other combinations.

c.) Tracking paired with target localisation or system commands.
(Wickens et al 1983)

The tracking task used a dual axis joystick to control a F/A 18 flight simulator. The target localisation task required subjects to move a cursor onto visually presented target, using a second joystick or spoken 'clock' commands. The command task required subject to enter a series of communication, navigation, or identification commands in response to prompts. Commands were either typed or spoken. Results show that the command task was aided by auditory prompting, but the localisation task was not. Further, they showed that performance decreased with shared visual inputs, but improved with shared manual responses.

d.) Threat evaluation paired with status checking
(Wickens et al 1984)

In the threat evaluation task, information relating to the position and heading of enemy aircraft is presented either via a CRT or by tones over headphones (the pitch and channel of the tones relating to spatial locations). Subjects had to decide if the aircraft represented a threat, and enter the level of threat ("low", "medium", or "high") either by pressing one of three corresponding keys or speech. The status checking task required subjects to speak or type a status check number. A visual or auditory display then informed the subject whether the part was o.k. or faulty. If the part was faulty, subject had to repeat the checks on components within that part. Again, responses were either typed or spoken. Results showed that performance on the status checking task decreased with auditory threat evaluation, and performance on the threat evaluation task decreased with visual status checking.

e.) Tracking paired with a verbal or spatial decision task
(Wickens and Liu 1988)

The tracking was the same as that used in (a.). The spatial decision task required subjects to assess the potential threat of enemy aircraft, in terms of movements along a vector display. Displays appeared every 12 seconds, and subjects had to compare the present display with the previous display to determine the bearing of the enemy aircraft. Subjects were required to evaluate the threat, and respond using buttons or speech. The verbal decision task required subjects to select weapons to destroy the target. Each weapon had a probability of success and a cost of use. Subjects had to balance success against cost in their selection. Decisions were forced paced at 12 seconds, and subjects responded using a keyboard or speech. Results showed that verbal decisions were faster and more accurate than spatial decisions. Further, spatial decisions were fastest using speech, and verbal decisions were fastest using keyboard.

Conclusions from Multiple Resource Theory Studies

Performance on a verbal secondary task, when paired with a spatial primary task, is best when auditory displays are with spoken responses (a.), (b.), (c.), (d.). Performance on a spatial secondary task, when paired with a spatial primary task, is best when manual responses to visual displays are used (c.), (d.). Performance on a verbal secondary task is better than that on a spatial secondary, when paired with a spatial primary task (e.). Performance on a verbal secondary task is best when responses a manual, and performance on a spatial secondary task are best when responses are spoken (e.) [c.f. all the other results]

From (a.) and (b.) we can conclude that pairing tasks which share the same central processing requirements will lead to performance decrements. From (c.), we can conclude that visual competition at input disrupts performance, especially if subjects are required to perform visual scanning with tracking. Further, spatial tasks show far more interference than combined verbal and spatial tasks (see also Baddeley and Lieberman, 1980). From (c.), it is also apparent that manual responses can be paired with other manual responses. This seems to contradict the initial hypotheses. However, because the secondary tasks required early, rather than late, processing, one would expect the bulk of interference to occur at input rather than output. This is consistent with the proposal the resource competition is related to the importance of that resource in performing a specific task (Navon and Gopher, 1979; Wickens, 1980).

From (d.), we can draw the same conclusions as from (a.) and (b.). However, the use of a tonal display for spatial location was found to be very difficult for some subjects, and highlights a problem of secondary task experiments. While some types of information can be displayed in either modality, depending on task requirements, other types of information can be presented in only one modality. Therefore, the necessity for assessment of information processing requirements of tasks, discussed in chapter eight, is supported by these studies. Before designing systems, it is not sufficient to rule out particular forms of interaction on the basis of "Aunt Sally" studies. In other words, rather than designing experiments to support multiple resource theory, one should design experiments which draw on the actual processing requirements of process control operators.

From (e.), we can draw the same conclusions as from (c.). The effect of memory load on performance is an important factor, and is further illustrated in a set of studies by Berman (1986). Subjects were presented with strings of five digits, either visually or auditorially. They repeated the string, either by typing or speaking. After 15 seconds, a second string was presented for a recognition test. Response time for the recognition test was significantly slower when auditory presentations were used, and when subjects had to speak the string, than in other conditions. Thus, speech appears to disrupt the short term storage of verbal information. This finding is further supported by Morris and Jones (1987). Baddeley (1987) proposes that speech disrupts storage in the phonological store of working memory.

Klapp and Netick (1988) presented subjects with two memory tasks. A "missing digit" task required subjects to attend to an eight digit string, and state which digit had not appeared from the set 0 - 9. A "probe digit" task required subjects to attend to an eight digit string, and then recall which digit followed a probe digit. The "missing digit" task only requires items to be retained until the missing digit is found, but the "probe digit" task requires digits to be retained in order. Thus, the latter places a heavier load on the subjects short term memory.

When the digits were presented, individually, subjects were asked to either say the digit or say "la la" (a form of articulatory suppression). Klapp and Netick (1988) found that articualtory suppression disrupted performance on the "probe digit" task, but had little effect on the "missing digit" task. However, the "missing digit" was disrupted by concurrent spatial processing tasks, while the "probe digit" task was not. In other words, when stimulus and response were kept constant, changes in processing demands produced different levels of interference (see also Greenwald, 1970 on 'ideomotor compatibility').

This study shows that the concept of multiple resources is limited. Both the "missing digit" and "probe digit" tasks would be defined as verbal tasks, and it is difficult to propose why a spatial task should disrupt the "missing digit" task. Possibly, subjects imagined the string of nine digits, and mentally crossed off each digit as it appeared. This could be defined as a form of spatial processing, albeit on verbal material.

These studies support the notion of 'C' as a limited attentional resource, employing either a verbal or spatial code of processing. Further, they show that compatibility can be obtained either across S-C-R, S-C, S-R, or C-R and between different aspects of C. This makes defining the processing codes employed difficult, if one wishes to progress beyond the simplistic dichotomy of verbal and spatial information. However, this is the only theory ,to date, which allows consideration of speech as a means of interaction with computers, and has received support from the studies. It allows hypotheses to be generated concerning operator performance when using ASR on different types of task. Before considering the viability of multiple resource theory to control room systems, we need to address some of the problems the theory has had to face.

Problems with Multiple Resource Theory

The studies reported from Wickens and his colleagues have been criticised for a number of methodological flaws. The most obvious was pointed out by Damos (1985; 1986). She noted that several of the studies employed a 'within subjects' experimental design. That is, the same subjects performed in different conditions. Such a design has been demonstrated to result in asymmetric transfer of learning from one condition to another (Poulton, 1982), with practice in one condition affecting performance in other conditions. However, the criticism of poor experimental design, need not affect the validity of multiple resource theory for three reasons. First and foremost, Wickens (1980) used a between subjects design, to arrive at similar findings. That is, when studies use acceptable designs, the results support the theory. Second, as Vidulich (1988) has pointed out, Damos (1985; 1986) uses pairs of discrete tasks for her study, rather than a continuous tracking task. This could allow subjects to schedule the demands of the tasks and so reduce resource competition. Thus, she may have been investigating a different aspect of multiple resource theory rather than competition between processing codes. Third, Damos (1985; 1986) does not offer an alternative explanation of either her findings or those of Wickens and his colleagues.

A second methodological flaw in Wickens' studies was noted by Hapeshi and Jones (1989b). They point out that, in general, the tracking tasks appear on one CRT, while the secondary tasks appear on a second CRT. While this could be argued to simulate the environment of the cockpit, it could also affect results. For instance, (c.) demonstrated that visual monitoring of feedback disrupts tracking performance. This could conceivably be due to the need to switch attention between CRTs rather than resource competition. That is, the results could arise from resource scheduling problems rather than resource competition. This hypothesis can be tested quite simply by using a single CRT for both displays.

Hapeshi and Jones (1989b) found that tracking was disrupted by the appearance of a visual prompt, which they suggest was due to competition for a single visual input channel. This was further supported by the fact that tracking is most disrupted by the appearance of the prompt, that is when two types of visual information appear on the screen. This finding can be accommodated within the multiple resource theory.

The study of Hapeshi and Jones (1989b) illustrates a number of problems with experiments using secondary tasks. Secondary tasks can disrupt primary task performance. This means that the results may not reflect performance on one or other of the tasks, but an interaction between the two. Consequently, changes in primary task performance can be overlooked, or results can be affected by other aspects of attentional resources, such as scheduling. Therefore, secondary task studies should employ a between subjects design, with the secondary task assessing resource capabilities required by the operator for the primary task, and both tasks appearing on a single screen. This will reduce the possibility of confounding results with extraneous variables.

Wickens (1984) multiple resource theory and stimulus response compatibility theory represent two sides of the same coin. From figures 14.2 and 14.3, it is clear that in some circumstances these approaches will yield different predictions (or different post hoc explanations). It is not clear how one should separate these two approaches, nor which of the two one should use for generating hypotheses. This ultimately leads to an theory which cannot be falsified, and which, therefore, is a dubious scientific validity. However, despite the objections discussed above, Wickens' theory is the only one capable of handling the psychological problems of speech technology.

Multiple Resource Theory in Process Control Operation

The studies reported from Wickens and his colleagues, above, were performed in relation to aviation psychology. They addressed issues of resource competition in cockpits of fighter aircraft, and used tracking tasks as the primary task. Tracking was performed manually. In the chapter one, we saw that process control operation rarely, if ever, uses manual tracking of plant or process behaviour. Therefore, one must be very careful in generalising Wickens results to the control room. There have been relatively few studies investigating the problems of multiple resources and ASR use in control rooms; a literature search only revealed one published study.

Wickens and Weingartner (1985) paired a process control monitoring task with either a verbal or a spatial secondary task. They found that monitoring was disrupted by the spatial secondary task, but not by the verbal secondary task. One can characterise monitoring as being primarily a visual spatial task, and multiple resource theory would predict an interference in the direction found by Wickens and Weingartner (see also Baddeley and Lieberman, 1980). However, monitoring is only one aspect of process control. The question remains to be asked, can ASR be used in conditions of high workload in process control rooms?

Study Eleven

From the discussion in chapter six, ASR is best suited to complex, verbal tasks, such as issuing commands. If ASR is used to control a simple process, one would anticipate performance to be disrupted more by a verbal reasoning task than a spatial one. If this is the case then using ASR for command and control can be beneficially paired with the visual spatial task of monitoring performance. Study eleven addresses this question.

20 employees of British Gas Midlands Research Centre (18 men and 2 women) volunteered to participate in this study. Experimental sessions lasted 45 minutes, and comprised three parts.

All subjects received an explanation of the process control simulation used in study eight (reported in chapter eleven). Following the explanation, subjects were randomly assigned to either a text or symbol feedback group. This follows on from studies seven and eight (chapter eleven) , and examines the role of feedback adjacent to plant items vs. type of feedback presentation. Subjects completed four sessions to control the plant, or as many sessions as it took to produce an output of over 90 units. As in study eight, subjects spoke the command to the experimenter, who translated it into keystrokes. The keystrokes produced feedback adjacent to the selected valve, and subjects had to respond using buttons as to the correctness of the feedback. Reaction time to feedback and output levels achieved were recorded, and compared using two way ANOVAs.

Following this initial session, all subjects were given an 8 item spatial reasoning test, followed by an 8 item verbal reasoning test. The spatial reasoning test required subjects to compare a perpendicular drawing of a stick man with a rotated version, and state whether they were the same or different (see the examples in figure 14.4).

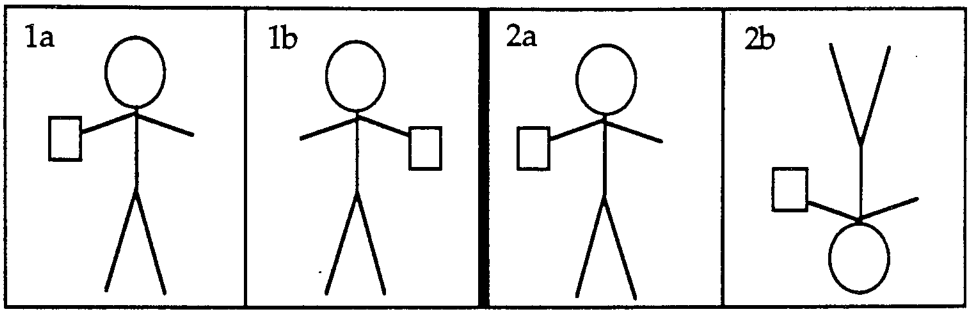


Figure 14.4: Examples of Items in Spatial Reasoning Test

The verbal reasoning test used was taken from Baddeley (1987). Subjects are presented with a pair of letters, prefixed with a statement which describes their relationship. For example,

A is before B.	AB	true/ false
A is not after B.	AB	true/ false
A is not before B.	AB	true/false

Subjects are required to state whether the statement is true or false. Subjects were allowed as much time as they liked to complete the tests. The subjects completed both reasoning tests within 20 seconds, on average. There were very few errors for the two tests: for the verbal reasoning test, there were 2 errors overall (out of a possible 160 items), and for the spatial reasoning test, there were 4 errors overall (out of a possible 160 items). These results suggest that the test were of comparable difficulty.

Following the reasoning test practice session, subjects performed the workload study. Subjects remained in the feedback group they were assigned to in part one, and further divided into either verbal or spatial secondary task groups. This gave a 2 x 2 experimental design, with 5 subjects in each group.

A window was included in the process plant display to show the secondary tasks. Each secondary task item remained on the screen until the subject answer it. Five seconds after the subject answered an item, another item appeared. Subjects responded using two buttons (yes/no). After a familiarisation session, subjects completed one experimental session. The experiment was limited to a single session for two reasons. First, in order to reduce the amount of time subjects were required to leave their work. Second, and more importantly, high workload situations are rare in control rooms (see Introduction). In high workload situations, operators often need to develop strategies and plans in response to novel events. Therefore, they do not have an opportunity to practice procedures or learn to time share in such situations. The use of a single session is intended to reproduce such conditions.

Reaction times to secondary tasks were recorded and compared using ANOVA. Note was also taken of the correctness of response and the number of items answered, together with overall output on process control.

Results

The results from the initial plant control stage of the experiment and the workload study are presented. The results section is divided into several parts, each dealing with a difference performance measure.

i.) Plant Control

Results from part one are divided into reaction time data and output level data. ANOVA tables show the levels of statistical difference between the two groups, and graphs are used to convey the overall differences in performance.

Reaction Time Data

A significant difference existed in performance across trials: F (3,51) = 10.471, $p<0.00001$. A Tukey test was performed, and showed that the difference was confined to the first two trials ($p<0.01$). There were no significant differences in performances over the second two trials. This is illustrated in figure 14.5. A significant level of interaction between type of feedback and trial is also apparent from figure 14.5: F (3,51) = 2.703, p<0.05. Investigation of the pooled error terms from this analysis revealed that, again the difference between groups lay in the first two trials, with textual feedback producing significantly faster reaction times than symbolic feedback. These points are illustrated in figure 14.6.

Source of variation	d.f.	Sum of Squares	F.	p.
Feedback	1	3178.884	3.354	0.085
error	17	16112.147		
Trial	3	4729.701	10.471	0.00001
Feedback / Trial	3	1221.065	2.703	0.05
errror	51	7679.097		

Figure 14.5: Table of ANOVA of reaction time data

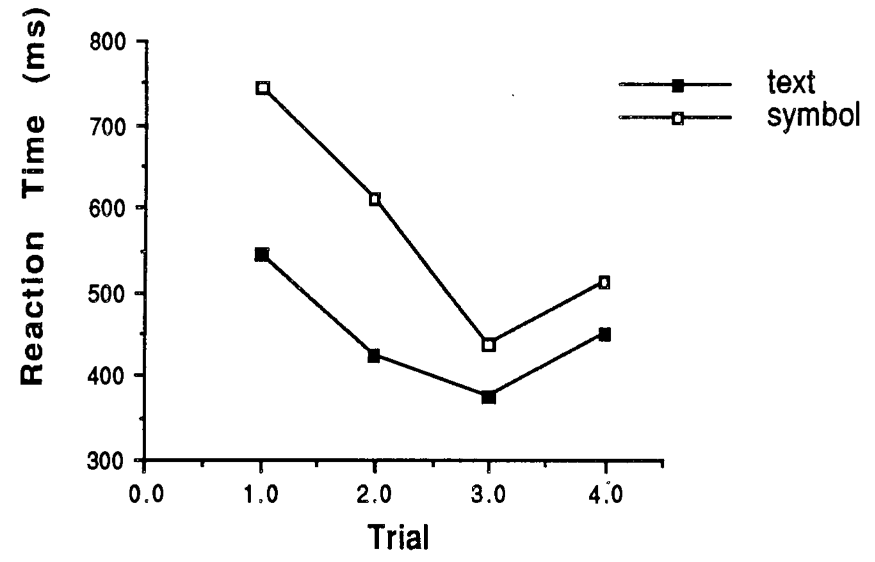


Figure 14.6: Graph of Reaction Times to Textual and Symbolic Feedback

Output Levels

Source of variation	d.f.	Sum of Squares	F.	p.
Feedback	1	1217.268	4.778	0.043
error	17	4331.413		
Trial	3	12017.2271	27.886	0.00001
Feedback / Trial	3	1029.822	2.390	0.079
errror	51	7326.035		

Figure 14.6: Table of ANOVA of output data

A significant difference was found across trials [F (3,51) = 27.886, p<0.00001], and between the two types of feedback [F (1,17) = 4.778, p<0.0431]. However, the differences between feedback types was confined to the first trial, as figure 14.7 shows. Both groups showed an improvement in performance, reaching criterion over the four trials.

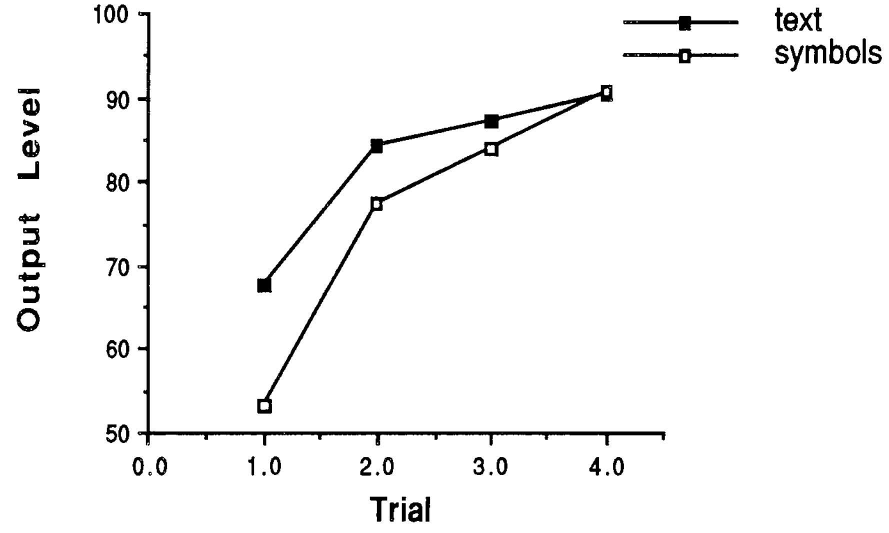


Figure 14.7: Graph of Output Level for Both Groups

ii.) Workload Study

The main performance measure for part three was the reaction time data for the secondary tasks. This was analysed using a two way ANOVA, of feedback type x secondary task. Analysis of output scores for the process plant task revealed that subjects were able to maintain process control performance. The output scores were not significantly different, and tended to a mean score of 94.5.

Source of variation	d.f.	Sum of Squares	F.	p.
Feedback	1	152.905	0.009	0.92
Sec. Task	1	427518.041	24.501	0.00001
Feedback / Sec. Task	1	3274.241	0.188	0.67
errror	16	7326.035		

Figure 14.7: Table of ANOVA for Secondary Task reaction times

There was no interaction between feedback type and secondary tasks, [F (1,16) = 0.188, p<0.6707]. There was, however, a significant difference between performance on the verbal and spatial secondary tasks: F (1,16) = 24.501, p<0.0001. Subjects responded more quickly to spatial items than to verbal items, as figure 14.8 shows.

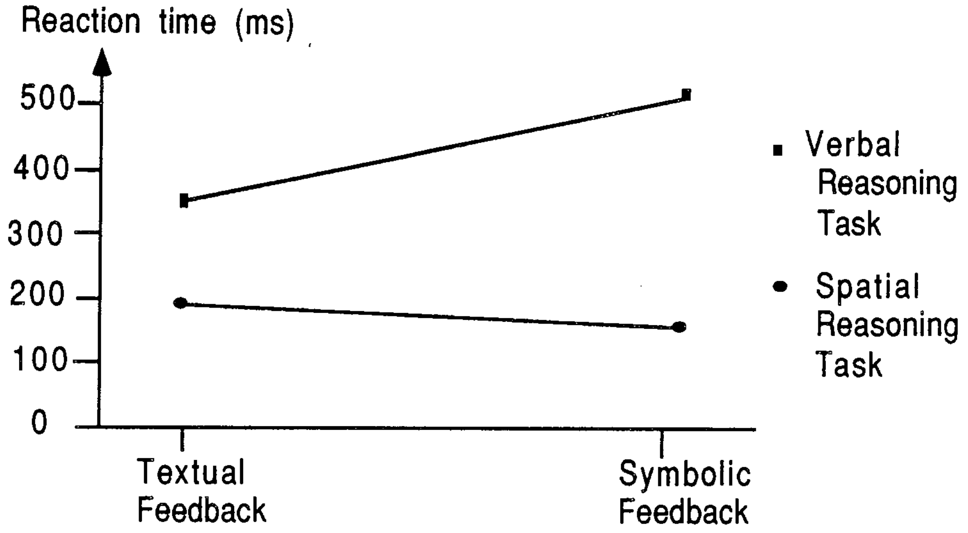


Figure 14.7: Reaction Time Data for Verbal and Spatial Reasoning Tasks

The difference in reaction times obviously means that subjects answer more items in the spatial reasoning task than in the verbal reasoning task (on average, 34.6 items on the spatial reasoning tasks, compared with 27.2 items on the verbal reasoning task). These data were not analysed statistically as they are derivatives of the primary performance measure, and support the main findings. Furthermore, subjects made more errors with the verbal reasoning task than with the spatial reasoning task (an error rate of 10.3% for the verbal reasoning task, compared with an error rate of 5.2% for the spatial reasoning task).

Discussion

The plant control task was intended to extend the studies reported in chapter eleven; specifically, to examine whether performance was related to the type of feedback used or the position of feedback. It showed that, in the initial trials, words were responded much more quickly to words than symbols. This is somewhat surprising given that the symbols were more distinctive than the words, and would be expected to be easier to discriminate. The symbol for "close" was a white arrow, pointing anticlockwise, in a red circle; the symbol for "open" was a white arrow, pointing clockwise, in a blue circle. In contrast, the words used were both in lower case (of approximately equal length), in white text on a black background. If subjects were required to make a decision concerning the meaning of the feedback, one would expect that the symbolic feedback would yield faster results, because discrimination between items is easier. From study seven, it was proposed that using ASR is, by definition, a verbal activity and would be best supported by verbal feedback. The results from part one support this finding.

The fact that differences in reaction times in later trials were negligible suggests that words and symbols could be used with equal efficacy as means of recognition feedback, providing they are placed adjacent to the object selected (see study eight). This means that if feedback is incorporated into the display relevant to the task being perfomed it will assist performance, and reduce the likelihood of user error (see study eight). This is true irrespective of the type of feedback provided. The fact that the symbols bore some resemblance to the valves may have affected initial performance. Many studies in psychology and ergonomics show that search and reaction times are increased if target and backgrounds are similar (e.g. Hitt, 1961). This leads to the recommendation that symbols be as distinct as possible from icons used in mimic displays. However, it is proposed that symbols can contain information in smaller areas than words, and will have a lower likelihood of overwriting the display. For these reasons, they are recommended for feedback concerning recognition performance. This would allow object selection and action specification to occur in one command utterance. The fact that ASR devices are prone to recognition error means that the user needs information as to what has been recognised. Such information is best presented in a text window below the display (see study eight).

The workload study examines the issue of feedback from the perspective of secondary task performance. We have already defined

ASR use as a verbal task. Multiple resource theory would predict that ASR use would be disrupted most by a verbal reasoning task. This is what part three found. However, research into the control of processes would lead one to assume that process control (see chapter one) is primarily a visual spatial task: the operator needs to consider the spatial location of system components and their physical relationships. In other words, process control çould represent a form of "cognitive tracking" behaviour, which would be disrupted by spatial reasoning tasks.

I would suggest that command and control is a higher level of control operation than system monitoring, and thus require more attention from the operator. Consequently, asking subjects to speak commands would employ an appreciable portion of verbal processing resources and result in low performance on a verbal reasoning task. The fact that performance on a spatial reasoning task was much better than on a verbal reasoning task supports this notion, and leads to the conclusion that process control can be performed using ASR. Operators will be able to perform command and control tasks in conjunction with spatial tasks, such as system monitoring.

The results of this study are obviously affected by the fact that the "ASR" device did not make any errors. In general, ASR devices misrecognise between 1% and 20% of words spoken. There is much that a system designer can do to improve accuracy, but errors are inevitable. This leads to two points for consideration of ASR in the control room. Operators need to be provided with an efficient means of error correction (chapter twleve), and ASR should be used for routine control operations, rather than in incidents.

Conclusions

The results from study eleven should be considered in the overall context of ASR system design for control room operation. They support, and extend the findings of studies seven and eight (chapter eleven). ASR will be best used in situations requiring complex verbal activity (see chapter seven). Using ASR for command and control is a verbal activity, which is best supported by verbal feedback. This feedback should be presented using text (see chapter eleven). However, use of ASR for system control requires two forms of feedback: one to inform the operator of the effects of his command (task feedback), and one to inform the operator of the validity of device recognition of his command (recognition feedback).

As Rosinski et al (1980) demonstrate, performance on a data entry task can be performed with equal success irrespective of the type or amount of feedback provided. However, correcting errors is most efficient if subjects are provided with a record of the data they have entered. This relates to the point made above. For process control operation, task feedback can be provided using symbols incorporated into the mimic display. This reduces operator error, and allows operators to maintain eye contact with the display. Conversely, recognition feedback requires a display of the recognised command, on a separate display, such as a text window.

Study eleven demonstrates that, in line with the predictions of multiple resource theory, the fact that ASR use is a verbal activity means that it can be paired with a spatial reasoning task without detriment to either task. Thus, using ASR for command and control will support visual scanning and monitoring of mimic displays. This develops the general argument in favour of ASR, presented in the Introduction and chapter three, that it allows 'eyes free' interaction with a computer. ASR allows the capacity for pairing spatial decision tasks with ASR use. The results of study one show that 'eyes free' use per se does not guarantee improved performance. Rather one needs to consider what tasks the operator will perform and how to restructure them to utilise ASR efficiently (see chapter eight).

CHAPTER FIFTEEN

CONCLUSIONS

This chapter summarises the major findings of the studies reported in this thesis. It is concluded that automatic speech recognition can offer benefits in specific, well defined areas.
Automatic speech recognition should be considered as an alternative to traditional media of human computer interaction, in situations involving a significant level of complex verbal activity, such as command and control.
Automatic speech recognition should be viewed in terms of other input media, and assessed accordingly. It should not be viewed as a panacea for all the ills of human computer interaction, nor as a means of over 'humanising' the interaction.
Recommendations are derived from the studies reported in this thesis, concerning the appropriate use of automatic speech recognition in the control room.

Introduction

This thesis has investigated many of the human factors issues surrounding the potential application of automatic speech recognition (ASR) to control room systems. From the discussion in chapter one, it was concluded that the research project should concentrate on the use of ASR in the context of human computer systems. This led to two important proposals. First, it defined a field of study, illustrated by figure 1, which comprised five main topics of human factors research. These topics were: training of the operator; the design of speech based interactions in control rooms; feedback for ASR use; error correction; and the use of ASR in conditions of high cognitive workload. Each of these topics was investigated in subsequent chapters, and the main conclusions of the studies will be reported below.

The second proposal which arose from defining the research project in terms of a human computer system was that ASR was to be regarded as a medium of human computer interaction rather than as a medium of communication. This point is often overlooked in the literature; ASR is too often regarded as a revolutionary form of human communication, when it should be regarded as a new means of inputting data into computers. This leads to a definition of speech based human computer interaction (HCI) which emphasises the actual tasks for which ASR is to be used, rather than the glib assumption that ASR will be 'natural' because it relies human speech. 'Naturalness' is not only a spurious concept, but also one which could result in the mismanagement of research projects. ASR is a novel medium of HCI which can offer benefits in a range of tasks and applications, in comparison with existing manually operated media. An aim of this thesis was to define the situations in which ASR could be beneficial in the control room.

Chapters three and four discuss the nature of speech and current approaches to ASR. I felt that it was important to describe how ASR worked, in order to appreciate its limitations and its benefits. This discussion concerns current techniques which have been implemented in commercially available devices, and thus does not consider the wealth on ongoing research which is improving the capability of ASR technology. While it is possible that some of the recommendations provided in this thesis will be superseded by advances in technology, I feel that the consideration of ASR system design from a human factors perspective will be relevant to the control room. Before discussing the findings of this thesis, it will be useful to provide an indication of the current state of the art systems upon which this research is based.

There seems to be a trade off in commercially available ASR devices in terms of the necessity for users to enrol devices and the size of available vocabulary. Devices which do not require user enrolment have small, predefined vocabularies. This means that the benefit of speaker independence will be offset by the number of available words, and by the fact that the vocabulary will be predefined. This will mean that it may be difficult to alter the vocabulary if the task changes. Therefore, it is recommended that control rooms use speaker dependent devices as they permit larger vocabularies and allow greater flexibility than speaker independent devices. Speaker dependent devices can handle vocabularies of between, on average, 64 to 1000 words. This might seem small compared with vocabulary of normal English speakers. However, from a review of current applications and an investigation of how people actually speak to ASR devices, it was argued that ASR systems will generally require vocabularies of less than 100 words. This is well within the capabilities of current devices. Finally, ASR devices can either recognise isolated words, in which the user is required to pause between each word spoken, or they can recognise connected words, in which the user is permitted to string words together to form phrases. It was proposed that in the control room, connected word ASR devices be used. However, ASR devices are unable to recognise conversational speech. This means that careful consideration needs to be given to how the user will speak to the device. While the current capability of ASR appears to be poor in comparison with human language use, it can be successfully applied in a number of situations. The comparison with human speech emphasises the erroneous assumption that ASR is a communication medium, rather than a medium of HCI. If one uses the latter definition, it will be clear that ASR should be considered in the context of other computer input media and not in the context of other speech users. From this, it appears necessary to compare ASR with other input media. However, this is not as easy a task as it may at first appear.

When to use ASR in the Control Room

ASR devices bear a similarity to function keyboards or touch screens, in that single commands can be substituted for a sequence of keystrokes on a qwerty keyboard. As one of the main aims of human factors is to design systems which reduce the possibility of human error and support human action, and as the majority of control room operators are not trained typists, it is preferable to reduce the amount of typing required of them. Any device which can reduce typing, and its attendant problems should be considered for control room operation. If the device can not only reduce problems of operator error but also improve their performance, then the device should be implemented.

Function keyboards are hard wired. This means that they are designated to perform specific functions in specific task domains. If the task domain alters, then the function keyboard needs to be replaced by one which matches the new domain. ASR provides more flexibility. The basic processes remain constant in different domains; only the software needs to be altered. Although it is easy to underestimate the problems involved in software modification, one can suggest that, in principle, this could be easier to do than rebuilding a function keyboard. ASR could be regarded as extensible function keyboard; speech commands represent several keystrokes.

Both function keyboards and touch screens are useful 'pointing devices', that is they support the action of selecting objects on displays. However, after an object has been selected the operator needs to use an additional input device to define what operation he wishes to perform on that object. ASR permits both 'pointing' and command entry. This means that, in principle, ASR offers a greater degree of flexibility in terms of operator activity than either function keyboards or touch screens. However, rather than assuming ASR will be the best device for all applications, it is necessary to define when to use ASR.

The review of applications presented in chapter six and the comparison of input media presented in chapter seven, led to the conclusion that ASR was best suited to tasks which can be defined as complex, verbal tasks. However, it was not clear whether the term 'complex' referred to the nature of the data used with ASR or the nature of the task. This issue was addressed by study one (chapter seven).

Conclusions from Study One

Study one compared the use of an isolated word ASR device with a function keyboard on a simulated process control task. An isolated word ASR device was used to provide an equitable performance with the function keyboard; one keystroke was assumed to equal a single spoken word. However, the performance of the subjects using ASR led to the conclusion that a direct comparison was somewhat misguided. The presence of an inherent time delay in the ASR condition confounded the results (which were based on time related measures). This led to the conclusion that comparison of input devices with different performance characteristics will not yield useful results (see also Carey, 1985).

Despite this problem of comparison, study one did produce three conclusions worthy of note. These concern the difference between male and female users of ASR; the relationship between task complexity and ASR use; and the nature of 'eyes free' ASR use.

1. Performance differences between male and female users of ASR

In terms of both recognition accuracy and task performance, male subjects achieved significantly better results than female subjects. This replicates a common problem for ASR use; women have trouble using speech recognition devices. It was assumed that this finding is directly related to the recognition algorithm employed by the Votan ASR device used in this study. Like many commercially available ASR devices, the Votan uses processing equivalent to filterbank analysis. Male speech generally produces lower frequencies than female speech. If the filterbank is set up to perform well using male speech, it may miss some of the speech information in the female speech. Thus, the problem does not lie in the female users but in the design of the system. Possibly this result could be eliminated if recognition algorithms were used which either did not rely solely of frequency information, or which were developed to work with the frequencies of female speech.

2. 'Complexity' in ASR use: a measure of task difficulty or message content ?

Process control operation has been defined as a complex task (see chapter two). From the definition of ASR being suited to complex, verbal tasks it could be assumed that ASR would be useful for the complexity of process control. However, in this study it is clear that the function keyboard produces better performance. If one examines the nature of commands the task requires, it is clear that the commands consist either of single words or two words. In terms of message content, this can be defined as a 'simple' task. This led to the conclusion that ASR will be suited to tasks in which the content of the message is both verbal and complex, such as in command and control situations which require the entry of detailed messages.

3. The nature of 'eyes free' ASR use.

It is common for writers to urge the use of ASR as an 'eyes free' medium of HCI. This means that in situations which require the operator to look away from the main keyboard or control console, ASR is proposed to be more beneficial than other media. However, study one shows that in some circumstances, 'eyes free' use may actually hinder performance. While the function keyboard structured subjects' activity, forcing them to enter commands using a set sequence, ASR allowed subjects to 'skip' between levels of the plant display. This resulted in a more erratic control performance from the subjects who used ASR.

Thus, rather than glibly recommending ASR for use in 'eyes free' situations, it is important to decide what the operators eyes are free for and what task they are performing. In other words, one needs to clearly define the task and the task domain in which ASR is to used. This was the subject of study two (chapter eight).

Conclusions from Study Two

Obviously it is not sufficient to base applications decisions on a list of *possible* advantages ASR might have over other media (see the list on page 108 for some of the advantages cited in the literature). Research should be conducted into the feasibility of using ASR for particular applications in particular task domains. EPRI (1986) proposed the use of a weighted selection procedure for application selection. However, this procedure was tailored to a specific domain. If one is to examine the general domain of control room operation for use of ASR, then one will require a general technique.

Study two demonstrated the use of hierarchical task analysis as a means of capturing task and domain information, in terms of control room operator activity. It was concluded that while several activities could be supported by the use of ASR, it was not believed that ASR would *significantly* enhance performance in the majority of these activities. However, the activity of issuing commands to remote sites over a computer communications network offered scope for application. The current system was believed to be slow and unwieldy. Further,the task required the construction and issuing of five word commands. This was proposed to represent a complex , verbal task and thus be suited to ASR use. The findings from study two provided the specifications for a speech based telecommand demonstration which was assessed in study three (chapter nine).

Conclusions from Study Three

While it is possible to ask operators and staff how desirable they believed ASR was for their operations, they find it difficult to accurately rate the technology until they have had experience of it. The aim behind study three was two fold:

i.) to provide operators with the opportunity to use ASR and then provide comments on their perceptions of its applicability;

ii.) to assess the performance of operators on using an ASR system.

It was found that after a short period of familiarisation, operators could issue 72% of the commands used within acceptable limits. This result was held to support the use of ASR in control room operations. Operators felt that it would significantly contribute to their work. However, while this second conclusion is very favourable it should be treated with some caution. The current system was problematic in a number of ways, and any improvement would have been received favourably. Also, ASR was a novel technology to the operators. The operators were sceptical before using it, but after use were very impressed. This could have created a 'novelty' effect which biased their opinions. Therefore, it is difficult to decide whether the operators' response to the demonstrator reflected the human factors system design which produced a better system than the one they were used to, or whether it reflected an improvement in performance due to the use of ASR. It was possible to conclude that ASR can be introduced into system design to produce a workable, efficient means of command entry.

Human Factors of ASR

Following the definition of probable applications of ASR in the control room, it is important to discuss and research the human factors problems this will introduce. From the discussion in chapter one, it will be seen that the main points which arise concern: user training, the style of speech based interaction with machines, error correction, device and system feedback, and the use of ASR in situations of high workload. These points are discussed with reference to the relevant studies in the section below.

Training

While the majority of computer input media require manual operations which are not necessarily common everyday actions, the use of ASR requires the use of speech. Thus it uses an overlearnt skill which has a number of presuppositions associated with it. New users of ASR need to be trained to use ASR efficiently, without encouraging them to engage in extraneous speech activity which could disrupt the devices performance.

It does not necessarily follow that because speech is 'natural' that speaking to computers will also be natural. Rather, because speech is an overlearnt skill, new users of ASR will have to adapt the way they speak to use ASR efficiently. It is possible to advise new users to speak 'consistently'. However, because they cannot introspect upon their speech, they will be unable to alter that part of the way they speak which governs consistency. This issue was studied in study ten (chapter thirteen).

Conclusions from Study Ten

It is proposed that, following the distinction drawn between declarative and procedural knowledge, in skill based activity, that speech requires procedural knowledge. From this, it is proposed that learning to use ASR will be best supported by the provision of some means of practice. A demonstration was argued to provide 'practice by proxy', and allowed new users to hear how they ought to speak. This would provide them with an appropriate model of the task of speaking to a computer, which they could copy. It was found that performance of subjects receiving a demonstration was superior to that of subjects receiving verbal instructions. Therefore, it is proposed that new users of ASR receive a demonstration of how to speak to an ASR device, by an experienced user of ASR, before they begin enrolment and use of the device.

Speech based Interaction with Computers in the Control Room

In chapter ten, it was proposed that the term "dialogue" requires an analogy to be drawn between human conversation and HCI which fails to capture to nature of HCI. A set of three studies was performed to investigate the difference between human - human communication and speech based HCI. In all three studies, a 'wizard of oz' experimental technique was used. This means that rather than using an ASR device, speech recognition is performed by the experimenter acting as a 'computer' in a separate room to the subject. This technique allows subjects to use whatever speech they wish to use. As the studies are examining the style of speech based HCI, the fact that subjects can choose the speech they feel to be most appropriate is an important consideration. Study four compared the performance of subjects issuing spoken commands to a 'computer' or to a human experimenter. Study five analysed the speech of subjects speaking to the 'computer', in order to define the style of speech people adopt in their speech based HCI. Finally, study six investigated the effect of different forms of feedback on subjects speech.

Conclusions from Study Four

Subjects who spoke to a human typist, used more words per command and a wider vocabulary than subjects who spoke to the computer. This suggests that people will restrict their speech in order to compensate for assumed limitations of computer ASR capabilities. Subjects talking to the 'computer' used short, succinct commands.

Study four also showed that people who spoke to the 'computer' issued more commands than those who spoke to the 'human'. This suggests that, because speech is the only channel of communication available to the 'computer' group, that the subjects speaking to the 'human' were able to maintain a dialogue using extralinguistic information, such as nodding or pointing. It also suggests that where the subjects in the 'human' group were able to make assumptions concerning what the experimenter could see (both subject and experimenter were looking at the same screen), in the 'computer' group, it was not clear what knowledge the 'computer' had. This meant that subjects assumed a greater role in controlling the process. Finally, study four showed that speaking to a 'computer' was not so much a conversation as a sequence of command exchanges.

Conclusions from Study Five

There are a number of conclusions which can be drawn from an analysis of the speech of subjects in this study. The first relates to the structure of the commands subjects used. It was anticipated that subjects would use a construction <verb><object>. However, only one of the subjects consistently used this construction, the other subjects simply named the object. This suggests that subjects are using ASR as pointing device in this context, that is they are using spoken commands to specify objects rather than issue complete commands. Further, there is an assumption that the presence of the <verb> is therefore redundant, presumably because the 'computer' 'knows' that in selecting an object, the subject also wishes a change in state to be performed on that object. This means that subjects assume that subject and 'computer' share knowledge about the language being used, which makes the use of 'extra' words, such as the explicit use of the <verb>, redundant.

Subjects use speech which they feel is most appropriate to the task and which is suited to the assumed capabilities of the 'computer'. As well as using shared information, subjects also tended to use use shared representations of that information. Thus, the type of feedback subjects used affected their choice of descriptors for plant units. Plant units could either be called by their colour name or by the letter which the colour begins with. There was a relationship between the use of word or letter and the area of the screen subjects used for feedback. If subjects used the operation log, they tended to use letters; if they used the graph, they tended to use words. However, analysis of 'reset' commands suggests that different types of feedback will be required to provide different types of system performance information.

Study five shows that, in speech based HCI, people will tend to use short, succinct phrases which are task specific. Errors are corrected by simply repeating the misrecognised command. It was proposed that speech based HCI is not conversational; subjects do not maintain a dialogue with the computer, as they would with a human (see study four). Human dialogue is a sequence of speech with exhibits both global and local coherence. In study five, the coherence is only local for the speech subjects use, but the process control task provides a global coherence. This means that the dynamic aspect of the task would not relate to an ongoing dialogue so much as a developing process requiring discrete control commands to be spoken at points in the process.

Conclusions from Study Six

This study showed that the more verbose the feedback provided to subjects, the fewer words subjects would use. This contradicts some previous research. However, it was noted that, in the process control simulation used in these studies, subjects rely on several areas of the screen for information. This means that feedback should be clear and unambiguous. The verbose feedback not only indicated what the 'computer' had recognised, but also suggested the possibility of negotiation. This type of feedback would contradict the proposal that speech based HCI is not conversational, and lead to a degree of confusion on the part of the subject. Therefore, variations in the feedback provided to users of ASR will affect their interpretation of the ongoing task, and 'extra' feedback may well be a hindrance to performance. Thus, rather than treating ASR use as a form of conversation, one should design speech based HCI which employs short, succinct , task specific phrases and which uses simple, task relevant feedback. The issues surrounding feedback for ASR was further investigated in studies seven and eight (chapter eleven).

Conclusions from Studies Seven and Eight

Chapter eleven discusses the topic of feedback, and proposes that a three point definition of feedback be used. This uses 'reactive' feedback, in which users receive feedback directly from the controls they are using; 'instrumental' feedback, in which users receive feedback immediately from the system they are controlling; and 'operational' feedback, in which users are required to make a decision from the feedback the receive. It was proposed that ASR will normally require 'operational' feedback, and that in the control room, this will be provided by visually displayed information. This information would be best presented by a string of text containing the words recognised, with each word appearing as it was recognised.

In process control diagrams, there is a great deal of information. It was suggested that the use of textual feedback might require too much screen space, and proposed that symbolic feedback be provided. Study seven compared the use of symbols and text on a simple verbal decision task. It was found that text produced more faster reaction time performance than symbols. This led to the conclusion that text would be preferable for verbal decision tasks, such as detecting and correcting errors.

While study seven examined the differences between types of feedback on a simple verbal task, this thesis has been investigating the use of ASR for performing process control tasks. Therefore, study eight compared the use of a text window with symbols adjacent to items in a process control display. It was found that, while overall process control performance was similar for the two types of feedback, there were significant differences in terms of user error. The text window produced an increase in the amount of feedback either ignored or misread. This means that subjects would not respond to commands, or accept commands which had been misrecognised. From this, it was proposed that feedback could be further described in terms of its function.

Feedback for task performance should be incorporated into the task display being performed. Feedback can be incorporated using symbols. However, study eleven (chapter fourteen) shows that the fact that feedback is incorporated into the diagram is more important than whether the feedback is presented using symbols or text. Symbolic feedback was recommended in that it would contain more information in less space than text.

Feedback for ASR performance should be presented using textual feedback, in a text window which can show the command that has been recognised. This will make error correction easy. If one relies solely on task performance feedback, it will be difficult to detect exactly where the error lies. Chapter twelve discusses the issues surrounding error correction.

Conclusions from Chapter Twelve

Chapter twelve reviews a number of errors which can arise from ASR use. It discusses substitution, insertion and rejection errors which can be made by the ASR device, and also the type of user errors which can arise. From the discussion of device errors, the chapter discusses automatic error correction. It was concluded that although automatic error correction is possible, the final decision concerning the validity of recognised speech should be left to the user. From this, a number of different approaches to user based error correction were discussed.

It was proposed that error correction should not require users to perform extratask actions, such as carrying out an error correction 'dialogue'. Rather error correction should be made implicit in the task. The form of error correction proposed in chapter twelve was developed from the findings of chapter ten (on speech based HCI), and chapter eleven (on incorporating feedback into the task). It is proposed that in command and control situations, operators need to first construct commands and then issue them to the plant or system. By dividing the task into two components, it is possible to introduce an element of system security. Operators will have to construct the command using speech, and then perform a manual action in order to 'send' the command. Command construction will follow from the use of short, succinct, task specific speech. Operators will first speak the command string and then correct any errors by repeating misrecognised words. This form of error correction capitalises upon the tendency of people to correct errors by repeating words, and also can be carried out after the initial command construction task, thus not interrupting command and control. Editing can be finessed by the use of vocabulary syntax and by eliminating misrecognised words from subsequent recognition passes.

Conclusions from Study Eleven

Study eleven (chapter fourteen) was divided into two parts. The first compared the use of symbols and words adjacent to objects in a process plant diagram, and was discussed above in the conclusions from studies seven and eight. Study eleven was primarily intended as an examination of operators' performance under high cognitive workload conditions. It was found that while subjects could perform a spatial secondary task without detriment, a verbal secondary task was affected. This led to the following conclusions, which support the findings of previous studies.

Speech affects performance on a verbal secondary task. This means that the use of speech recognition is in itself a verbal task, and should not be paired with other verbal tasks, such as communication or data entry. This also supports the conclusion from study ten, that ASR use constitutes a verbal skill.

Speech does not affect performance on a spatial secondary task. This means that speech recognition can be paired with a spatial task, such as monitoring plant conditions. This supports the finding from study one, that 'eyes free' use *per se* will not guarantee improved performance, but one needs to consider the tasks for which ASR is being used together with the tasks being performed in conjunction with ASR use.

Proposals for Future Research

This thesis has not only demonstrated that ASR is viable for command and control tasks, but also made recommendations concerning human factors system design.

Future research can follow a number of avenues, the most pressing need is to develop and assess a large scale ASR system. The process control simulations reported in the thesis have been realistic, but small scale. They require vocabularies of 50 words rather than 1000. Increases in vocabulary size bring with them decreases in recognition performance.

A second line of research can be conducted into intelligent error correction techniques. ASR devices which are currently available serve as 'front ends' to more sophisticated language processors. Research is currently being performed by a number of institutions to improve the 'front ends' speech processing capability. This lies outside the remit of human factors research, but will obviously affect system performance.

This thesis has defined a field of human factors research and investigated the topics contained in this field. A number of conclusions have been drawn with reference to the human factors design of ASR systems. There are several topics that have arisen from the research which could be studied further.

Human factors research can investigate the problems of stress of human speech. This might be a problem for ASR in the control room. However, stress can be assumed to relate to non routine situations, which are, by definition, rare in control room operation. Therefore, it is proposed that ASR be used in routine operations.

A second human factors project could investigate the use of ASR for problem solving activities. For instance, while it is possible to drive a car and hold a conversation in light traffic, as traffic becomes more dense people concentrate on the driving task at the expense of the conversation. Does the use of ASR in problem solving situations affect performance? Study eleven showed that ASR use could be paired with specific *types* of secondary task. It would be interesting to investigate whether ASR could be used for complex problem solving exercises, such as fault finding using an expert system.

Human factors research can be addressed to the design of vocabularies. It was suggested that, in chapter ten, vocabulary design arises from the task domain and the style of speech based HCI. While the studies reported in chapter ten define the style of speech base HCI for process control tasks, it did not address the selection of vocabulary items for operation. In the other studies reported vocabulary arose from the task domain investigated. There might be a relationship between the ease with which a device can recognise words and the 'habitability' of the language used for operators.

In summary, this thesis has proposed that ASR can be used for command and control operation in the control room. This defines a limited subset of control room activities. Speech based HCI will comprise of a short, succinct, task specific style of speech. ASR use is a verbal skill which requires training by demonstration, and which can be paired with spatial secondary tasks, such as monitoring. The use of ASR should be carefully considered to examine as many of the factors in a specific task domain as possible. This thesis recommended the use of HTA. Finally, applications decisions should be based around demonstration systems which can be assessed by prospective users. This can then lead into rapid prototyping, based around the human factors system design proposals in this thesis.

Pages 322 - MISSING

REFERENCES

Ainsworth, L. and Whitfield, D. (1983) **The Use of Verbal Reports for Analysing Power Station Control Skills** Birmingham: Aston University Applied Psychology Department Report No. 114

Ainsworth, W.A. (1988) **Speech Recognition by Machine** London: Peter Peregrinus

Allengry, P. (1987) The analysis of knowledge representation of nuclear power plant control room operators In Ed. H.J. Bullinger and B. Shackel **Interact'87** Amsterdam: Elsevier

Allerhand, M. (1987) **Knowledge Based Speech Pattern Recognition** London: Kogan Page

Allport, D.A. (1980) Attention and perofmance In Ed. G. Claxton **Cognitive Psychology: New Directions** London: Routledge Kegan and Paul

Anderson, J.R. (1980) **Cognitive Psychology and its Implications** San Francisco: W.H. Freeman and Co.

Anderson, R.P. and Gill, K.D. (1987) Applications of voice recognition techniques in the automotive industry **Proceedings of International SpeechTech'87** New York: Media Dimensions 9 - 11

Annett, J. (1989) Skills In Ed. A.M. Colman and J.G. Beaumont **Psychology Survey 7** Leicester: The British Psychological Society

Aretz, A.J. (1983) A comparison of manual and vocal response modes for the control of aircraft systems **Proceedings of 27th. Annual Meeting of the Human Factors Society** 97-101

Atal, B.S. (1985) Linear predictive coding of speech In Ed. F. Fallside and W.A. Woods **Computer Speech Processing** Englewood Cliffs, NJ: Prentice Hall

Austin, J.L. (1962) **How to do Things with Words** Oxford: Clarendon Press

Backer, K.A. (1984) Computers that listen are longer just talk! **Proceedings SpeechTech'84** New York: Media Dimensions

Baddeley, A.D. (1983) Working memory **Philosophical Transactions of the Royal Society of London B302** 311-324

Baddeley, A,D. (1986) **Working Memory** Oxford: Clarendon Press

Baddeley, A.D. and Hitch, G.J. (1974) Working Memory In Ed. G. Bower **Recent Advances in Learning and Motivation vol. VIII** New York: Academic Press

Baddeley, A.D. and Lieberman, K. (1980) Spatial working memory In Ed. R. Nickerson **Attention and Performance vol. VIII** Hillsdale, NJ: Erlbaum

Bahl, L.R., Cole, A.G., Jelinek, F., Mercer, R.L., Nadas, A., Nahamod, D., and Picheny, M.A. (1983) Recognition of isolated word sentences from a 5.000 word vocabulary office correspondance tasks **Proceedings International Conference IEEE ASSP** 1065-1067

Bailey, P. (1984) Speech communication: the problem and some solutions In Ed. A. Monk **Fundamentals of Human Computer Interaction** London: Academic Press

Bainbridge, L. (1974) Analysis of verbal protocols from a process control task In Ed. E. Edwards and F. P. Lees **The Human Operator in Process Control** London: Taylor and Francis

Bainbridge, L. (1981) Mathematical equations or processing routines? In Ed. J. Rasmussen and W.B. Rouse **Human Detection and Diagnosis of System Failures** New York: Plenum Press

Bainbridge, L. (1984) Diagnostic skill in process operation **Proceedings International Conference on Occupational Ergonomics** 7-9 May, Toronto

Bainbridge, L. (1987) The ironies of automation In Ed. J. Rasmussen et al **New Technology and Human Error**
Chichester: John Wiley

Bainbridge, L. (1988) Types of representation
In Ed. L.P. Goodstein et al **Tasks, Errors and Mental Models** London: Taylor and Francis

Baker, J.M. (1975) The DRAGON system - an overview
IEEE Transactions ASSP - 23 24-29

Baker, J.M. (1981) How to achieve recognition: a tutorial status report on automatic speech recognition **Speech Technology 1** (1) 30-43

Baker, J.M. and Pinto, D.F. (1986) Using commercially available adaptation across different speakers to achieve high recognition performance **Proceedings SpeechTech'86** 47-51

Balzer, W.K., Doherty, M.E., and O'Connor, R. (1989) Effects of cognitive feedback on performance **Psychological Bulletin 106** (3) 410-433

Barber, P. and Legge, D. (1976) **Perception and Information** London: Methuen

Barnard, P. (1974) **Unpublished Corpus of Directory Enquiry Conversations** Cambridge: MRC APU

Barnard, P. (1985) Interacting cognitive subsystems: a psycholinguistic approach to short term memory In Ed. A. W. Ellis **Progress in the Psychology of Language vol. II** Hillsdale, NJ: Erlbaum

Barnard, P. and Marcel, T. (1984) Representation and understanding: in the use of symbols and pictograms In Ed. R. Easterby and H. Zwaga **Information Design** Chichester: Wiley

Bartlett, F. C. (1932) **Remembering: a Study in Experimental and Social Psychology** Cambridge: Cambridge University Press

Batchellor, M.P. (1981) **Investigations of parameters affecting voice recognition in C3 systems** Monterey, CA: Naval Postgraduate School Unpublished Masters Thesis

Bauerschmidt, D.K. and LaPorte, H.R. (1976)Techniques for display/control system integration in supervisory and monitoring systems In Ed. T.B. Sheridan and G. Johannsen **Monitoring Behaviour and Supervisory Control** New York: Plenum Press

Berman, J.V.F. (1986) **Speech Recognition in High Performance Aircraft: Some Human Factors Considerations** London: Institute of Aviation Medicine Report No. 646

Berman, J.V.F. (1984) Speech technology in a high workload environment **Proceedings 1st. International Conference on Speech Technology** 69-76

Berry, D.C. and Broadbent, D.E. (1984) On the relationship between task performance and verbalizable knowledge **Quarterly Journal of Experimental Psychology** 36A 209-231

Berry, D. C. and Broadbent, D.A. (1988) Interactive tasks and the implicit - explicit distinction **British Journal of Psychology** 79 (2) 251-272 ,

Bierschwale, J.M., Sampaio, C.E., Stuart, M.A. and Smith, R.L. (1989) Speech versus manual control of camera functions during a telerobotic task **Proceedings Human Factors Society 33rd.Annual Meeting** 134-138

Bloc, L. (1978) **Speech Communication with Computers** London: MacMillan

Bobrow, D.G. and Klatt, D.H. (1968) A limited speech recognition system **Proceedings AFIPS Joint Conference** Washington, DC: Thompson 305-318

Boden, M. A. (1987) **Artificial Intelligence and Natural Man** Cambridge, Mass: MIT Press

Boles, D. and Wickens, C.D. (1987) Display formatting in information integration and nonintegration task. **Human Factors** 29 395-406

Bowles, R.L., Damper, R.I., and Lucas, S.M. (1988) Combining evidence from separate speech recognition processes **Proceedings Speech '88 - 7th. FASE Symposium** 669-675

Bond, Z.S., Moore, T.J. and Gable, B. (1986) Some phonetic characteristics of speech produced in noise **Journal of the Acoustical Society of America**

Bond, Z.S., Moore, T.J. and Anderson, T.R. (1987) The effects of high sustained acceleration on the acoustic phonetic structure of speech: a preliminary investigation **Journal of the American Voice I/O Society** 4 1-19

Brainard, R.W., Irby,T.S., Fitts, P.M. and Alluisi, E.A. (1962) Some variables influnecing the rate of gain of information **Journal of Experimental Psychology** 63 105-110

Brandau, R.J. (1982) Performance interactions with voice interactive systems: implementations of task complexity In Ed. D.S. Pallett **Proceedings of the Workshop on Standardisation for Speech I/O Technology**

Bransford, J.D. and Franks, J.J. (1971) The abstraction of linguistic ideas **Cognitive Psychology** 3 331-350

Brehmer, B. (1980) In one word: not from experience **Acta Psychologica** 45 223-241

Bristow, E. (1986) **Electronic Speech Recognition:Techniques, Technology, and Applications** London: Collins

Broadbent, D.E. (1958) **Perception and Communication** Oxford: Pergamon

Broadbent, D.E. (1977) Levels, hierarchies, and the locus of control **Quarterly Journal of Experimental Psychology** 29 181-201

Broadbent, D.E., FitzGerald, P. and Broadbent, M.H.P. (1986) Implicit and explicit knowledge in the control of complex systems **British Journal of Psychology** 77 33-50

Brown, G. (1984) Linguistic and situational context in a model of task oriented dialogue In Ed. L. Vaina and J. Hintikka **Cognitive Constraints on Communication: Representations and Processes** Dordrecht : Reidel

Brown, N.R. and Vosburgh, A.M. (1989) Evaluating the accuracy of a large vocabulary speech recognition system **Proceedings Human Factors Society 33rd. Annual Meeting** 296-300

Bruce, P.C. (1987) Engineering an intelligent voice dialogue controller **British Telecom Technology Journal** 5 (4) 61-69

Bruce, V. and Valentine, T. (1985) Identity priming in the recognition of familiar faces **British Journal of Psychology 76** 373-383

Bussak, G. (1983) Voice leaders speak out **Speech Technology 1** (4) 55-68

Byblow, W.D. (1990) Effects of redundancy in the comparison of speech and pictorial displays in the cockpit environment **Applied Ergonomics 21** (2) 121-128

Byford, R.G. (1987) Voice on the factory floor **Proceedings International SpeechTech'87** New York: Media Dimensions 1-6

Carbonnell, J. C. and Hayes, P.J. (1983) Recovery strategies for parsing extragrammatical language **American Journal of Computational Linguistics 9** 123-146

Card, S.K., Moran, T.P., and Newell, A. (1983) **The Psychology of Human Computer Interaction** Hillsdale, N.J.: Lawrence Erlbaum

Carey, M.S. (1985) **Selection Criteria for Input Devices: Interim Report** Aston University, Birmingham: Applied Psychology Division

Carney, T.F. (1972) **Content Analysis: a Technique for Systematic Inference from Communication** London: Basford

Carroll, J.M. (1984) Minimalist design for active users In Ed. B. Shackel **Interact '84** Amsterdam: Elsevier

Carter, R. C. and Cahill, M. C. (1979) Regression models of search time for colour coded information displays **Human Factors 21** 293-302

Carter, K.E.P, Newell, A.F. and Arnott, J. (1988) Studies using a simulated listening typewriter **Proceedings Speech'88: 7th. FASE Symposium** 1289-1296

Casali, S.P., Dryden, R.D. and Williges, B.H. (1988) The effects of recognition accuracy and vocabulary size of a speech recognition system on task performance and user acceptance **Proceedings Human Factors Society 32nd. Annual Meeting** 232-236

Cassford, G.E. (1988) Applications of speech technology in the automotive industry **Proceedings IEE Colloquium on Speech Processing** Digest no.: 1988/11

Cashdan, A. and Jordan, M. (1987) **Studies in Communication** Oxford: Basil Blackwell

Cegrell, T. (1986) **Power System Control Technology** London: Prentice Hall

Chapanis, A. (1975) Interactive human communication **Scientific American 232** 36-42

Chapanis, A., Garner, W.R. and Morgan, C.T. (1949) **Applied Experimental Psychology** New York: Wiley

Chapanis, A., Parrish, R.N., Ochsman, R.B. and Weeks, G.D. (1977) Studies in interactive communication: II. The effects of four modes on the linguistic performance of teams during cooperative problem solving **Human Factors** 19 (2)

Cherry, E.C. (1953) Some experiments on the recognition of speech, with one and two ears **Journal of the Acoustical Society of America** 25 975-979

Chistovich, L.A. (1979) Auditory processing of speech **Proceedings 9th. International Congress of Phonetic Sciences** 83-91

Chomsky, N. and Halle, M. (1968) **The Sound Pattern of English** New York: Harper and Row

Chomsky, N. and Miller, G.A. (1963) Introduction to the formal analysis of languages In Ed. R.D. Luce et al. **Handbook of Mathematical Psychology Vol. II** New York: Wiley

Christ, R.E. (1975) Review and analysis of colour coding research for visual displays **Human Factors** 17 542-570

Christ, R.E. and Corso, G.M. (1983) The effects of extended practice on the evaluation of visual display codes **Human Factors** 25 71-84

Clark, H.H. and Clark, E.V. (1977) **Psychology and Language: an Introduction to Psycholinguistics** New York: Harcourt Brace and Jovanovich

Clarke, R. and Morton, J. (1983) Cross modalitiy facilitation in tachistoscopic word recognition **Quarterly Journal of Experimental Psychology** 35A 79-96

Claxton, G. (1980) **Cognitive Psychology: New Directions** London: Routledge Kegan and Paul

Cochran, D.J., Riley, M.W. and Stewart, L.A. (1980) An evaluation of the strengths, weaknesses and uses of voice input devices **Proceedings Human Factors Society 24th.Annual Meeting** Santa Monica: Human Factors Soc 190-194

Cohen, N.J. and Squire, L.R. (1980) Preserved learning and retention of pattern analysing skill in amnesia: dissociation of knowing how and knowing that **Science** 210 207-210

Cole, R.A., Rudnicky, A.I., Zue, V.W. and Reddy, D.R. (1980) Speech as patterns on paper In Ed. R.A. Cole **Perception and Production of Fluent Speech** Hillsdale, N.J.: Lawrence Erlbaum

Connolly , D.W. (1979) **Voice Data Entry in Air Traffic Control** Atlantic City, NJ: Federal Aviation Administration

Conrad, R. (1960) Acoustic confusion in immediate memory **British Journal of Psychology** 55 75-84

Conrad, R. and Hull, A.J. (1964) Information, acoustic confusion and memory span **British Journal of Psychology** 55 429-432

Cooke, J.E. (1965) **Human Decisions in the Control of a Slow Response System** Unpublished PhD Thesis cited in L.Bainbridge (1981)

Cookson, S. (1988) Final evaluation of VODIS **Proceedings Speech'88: 7th. FASE Symposium** 1311-1320

Cotton, J. C. (1981) **A Classification System for Speech Recognition Technology** Working Paper, DoD cited in McCauley (1984)

Cotton, J. C. and McCauley, M.E. (1983) **Voice Technology Design Guidelines for Navy Training Systems** Orlando: Naval Training Equipment Centre Report No. NAVTRAQUIPEEN 80.C.0057.1

Cotton, J.C., McCauley, M.E., North, R.A. and Streib, M.I. (1983) **Development of Speech Input/Output for Tactical AircraftDayton,** Oh: AFWAL - TR - 83 - 3073, 1973 (AD - A136 485)

Coury, B.G. and Pietras, C.M. (1989) Alphanumeric and graphic displays for dynamic process control monitoring **Ergonomics** 32 (11) 1373-1389

Cox, S.J. (1988) Hidden markov models for automatic speech recognition: theory and application **British Telecom Technology Journal** 6 (2) 105-115

Craft, A.M. (1982) Human factors and ASR: a challenge for us all **Proceedings Voice Data Entry Systems Applications Conference**

Craik, F.I.M. and Lockhart, R.S. (1972) Levels of processing: a framework for memory research **Journal of Verbal Learning and Verbal Behaviour** 11 671-684

Craik, K.J.W. (1943) **The Nature of Explanation** Cambridge: Cambridge University Press

Craik, K.J.W. (1947) Theory of the human operator in control systems **British Journal of Psychology** 39 56-61

Crossman, E.R.F.W. (1956) The information capacity of the human operator in symbolic and nonsymbolic control processes **Information Theory andthe Human Operator** Min. of Supply pub.: WR/D2/56

Damos, D.L. (1985) The effect of asymmetric transfer and speech technology on dual task performance **Human Factors** 27 409-421

Damos, D.L. (1986) The effect of using voice generation and recognition systems on the performance of dual tasks **Ergonomics** 29 1359-1370

Damper, R.I. (1984) Voice input aids for the physically disabled **International Journal of Man Machine Studies** 21 542-553

Damper, R.I. (1988) Practical experiences with speech data entry In Ed. E.D. Megaw **Contemporary Ergonomics 1988** London: Taylor and Francis 92-97

Damper, R.I., Lambourne, A.D. and Guy, D.P. (1985)Speech input as an adjunct to keyboard entry in television subtitling In Ed.B. Shackel **Interact'84** Amsterdam: Elsevier

Damper, R.I. and MacDonald, S.L. (1984) Template adaptation in speech recognition **Proceedings of the Institute of Acoustics** 6 (4) 293-300

Danis, C.M. (1989a) Goats to sheep: can recognition rate be improved for poor Tangora speakers? **Proceedings of Speech and Natural Language Workshop** San Mateo, CA: Morgan Kaufman 145-150

Danis, C.M. (1989b) Developing successful speakers for an automatic speech recognition system **Proceedings Human Factors Society 33rd. Annual Meeting** 301-304

Davis, K.H., Biddulph, R. and Balashek, J. (1952) Automatic recognition of spoken digits **Journal of the Acoustical Society of America** 24 637-645

de Kleer, J. and Brown, J.S. (1983) Assumptions and ambiguities in mechanistic mental models In Ed. D. Gentner and A.L. Stevens **Mental Models** Hillsdale, NJ: Erlbaum

Dell, G.S. (1986) A spreading activation theory of retrieval in sentence production **Psychological Review** 93 283-321

Denes, P. and Matthews, M.V. (1960) Spoken digit recognition using t ime frequency pattern matching **Journal of the Acoustical Society of America 32** 1450-1455

Denes, P.B. and Pinson, E.L. (1963) **The Speech Chain: the Physics and Biology of Spoken Language** Garden City, N.Y.: Anchor Press

Department of Trade and Industry (1988) **Speech and Language Technology - Strategies for Research and Development Support** DTI: Speech and Language Technology Club

Dersch, W.C. (1962) Shoebox - a voice responsive machine **Datamation 8** 47-50

Deutsch, J.A. and Deutsch, D. (1963) Attention: some theoretical considerations **Psychological Review 70** 80-90

Devoe, D.B. (1967) Alternatives to handprinting in the manual entry of data **IEEE Transactions HFE - 8** 21-31

Dewer, R.E. and Swanson, H.A. (1972) Recognition of traffic control signs **Highway Research Record 414** 16-23

Dewer, R.E., Ells, J.G., and Mundy, G. (1976) Reaction time as an index of traffic sign perception **Human Factors 18** 381-392

Diaper, D. and Shelton, T. (1989) Natural language requirements for expert system naive users In Ed. J. Peckham **Recent Developments and Applications of Natural Language Processing** London: Kogan Page

Dillon. R.F., Edey, J.D., and Tombaugh, J.W. (1990) Measuring the true cost of command selection: techniques and results **CHI'90 Proceedings** 19-25

Dixon, N.R. and Martin, T.B. (1979) **Automatic Speech and Speaker Recognition** New York: IEEE Press

Doddington, G.R. (1980) Whither speech recognition? In Ed. W.A. Lea **Trends in Speech Recognition** Englewood Cliffs, NJ: Prentice Hall

Doddington, G.R.and Schalk, T.B. (1981)Speech recognition: turning theory to practice **IEEE Spectrum 18** (9) 26-32

Dreizen, F. (1987) An alternative way to handle errors: the sieve method for error detection and correction **Proceedings International Speech Tech '87** New York: Media Dimensions 192- 195

Dreyfus Graf, J. (1949) Sonograph and sound mechanics **Journal of the Acoustical Society of America 22** 731-739

Dudley, H. and Balashek, S. (1958) Auotmatic recognition of phonetic patterns in speech **Journal of the Acoustical Society of America 30** 721-739

Easterby, R. and Zwaga, H. (1984) **Information Design** Chichester: Wiley

Edwards, E. (1976) Preview of Process Control Section In Ed. T.B. Sheridan and G. Johannsen **Montoring Behaviour and Supervisory Control** New York: Plenum Press

Edwards, E. and Lees, F.P. (1974)**The Human Operator in Process Control** London: Taylor and Francis

Egeth, H.E. (1966) Parallel vs. serial processes in multidimensional stimulus discrimination **Perception and Psychophysics 13** 394-402

Ehrich, R.G. and Williges, R.C. (1986) **Human Computer Dialogue Design** Amsterdam: Elsevier

Ells, J.G. and Dewer, R.E. (1979) Rapid comprehension of verbal and symbolic traffic sign messages **Human Factors 21** 161-168

EPRI (1986) **Speech Recognition and Synthesis for Electric Utility Dispatch Control Centres** EPRI Report EL-4481 Project Report 2473-1

Erman, L.D. and Lesser, V.R. (1980) The HEARSAY- II speech understanding system In Ed. W.A. Lea **Trends in Speech Recognition** New York: Prentice Hall

Eysenck, M. (1984) **A Handbook of Cognitive Psychology** Hillsdale, NJ: Erlbaum

Fallside, F. and Woods, W.A. (1985) **Computer Speech Processing** Englewood Cliifs, N.J.: Prentice Hall

Fant, G. (1960) **Acoustic Theory of Speech Production** The Hague: Mouton

Falzon, P. (1984) The analysis and understanding of an operative l anguage In Ed. B. Shackel **Interact'84** Amsterdam: Elsevier 237-241

Fielding, G. and Hartley, P. (1987) The telephone: a neglected medium In Ed. A. Cashdan and M. Jordan **Studies in Communication** Oxford: Basil Blackwell

Fisher, M. (1986) Voice control for the disabled In Ed. G. Bristow **Electronic Speech Recognition** London: Collins

Fitts, P.M. (1954) The information capacity of the human motor system in controlling the amplitude of movement **Journal of Experimental Psychology** 47 381-391

Fitts, P.M. (1964) Perceptual motor skill learning In Ed. A.W. Metton. **Categories of Human Learning** New York: Academic Press

Fitts, P.M. and Posner, M.A. (1967) **Human Performance** Pacific Pallisades,CA: Brooks/Cole

Flanagan,J.L. (1972) **Speech Analysis, Synthesis and Perception** Berlin: Springer Verlag

Flanagan, J.L. (1976) Computers that talk and listen: man machine communication by voice **Proceedings of the IEEE 64 (4)** 405-415

Fleishman, E.A. and Quaintance, M.K. (1984) **Taxonomies of Human Performance** San Diego, CA: Academic Press

Flowers, J.B. (1916) The true nature of speech **Proceedings of the American Institute for Electrical Engineering** (09/01/16) 213-231

Ford, W.R., Weeks, G.D. and Chapanis, A. (1980) The effect of self imposed brevity on the structure of dyadic communication **Journal of Psychology 104** 87-103

Forren, M.G. and Mitchell, C.M. (1986) Multimodal interface for supervisory control **Proceedings Human Factors 30th. Annual Meeting** 317-321

Forster, K.I. (1970) Visual perception of rapidly presented word sequences of varying complexity **Perception and Psychophysics 8** 215-221

Frankish, C.R. and Jones, D.M. (1987) Parcel sorting by speech recognition: a case study in vocabulary design **Proceedings of the European Conference on Speech Technology** 197-201

Frankish, C.R., Jones, D. and Hapeshi, K. (1990) Maintaining recognition accuracy during data entry tasks using speech input In Ed. E.J. Lovesey **Contemporary Ergonomics 1990** .London: Taylor and Francis 445-453

Frauenfelder, U.H. and Tyler, L.K. (1987) **Spoken Word Recognition** Cambridge, Mass.: M.I.T. Press

Fromkin, V.A. (1980) **Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen and Hand** New York: Academic Press

Fry, D.B. (1979) **The Physics of Speech** Cambridge: Cambridge University Press

Fry, D.B. and Denes, P. (1958) The solution of some fundamental problems in mechanical speech recognition **Language and Speech 1** 35-58

Fujisaki, H., Hirose, K., Udagawa, H. and Kanedera, N. (1986) A new approach to continuous speech recognition based on consideration of human processes of speech perception **Proceedings IEEE Conference ASSP 86** 1959-1962

Gaines, B.R. and Shaw, M.L.G. (1983) Dialogue engineering In Ed. M.E. Sime and J. Coombs **Designing for Human Computer Communication** London: Academic Press

Garnham, A., Oakhill, J.V. and Johnson Laird, P.N. (1981) Slips of the tongue in the London - Lund corpus of spontaneous conversation **Linguistics 19** 805-817

Garrett, M.F. (1976) Syntactic processes in sentence production In Ed. R.J. Wales and E. Walker **New Approaches to Language Mechnaisms** Amsterdam: North Holland

Garrett, M.F. (1984) The organisation of processing structures for language production: applications to aphasic speech In Ed. D. Caplan et al **Biological Perspectives on Language** Cambridge, Mass: MIT Press

Gentner, D. and Stevens, A.L. (1983) **Mental Models** Hillsdale, N.J.: Erlbaum

Gerald, J.A. (1984) A voice response system for general aviation **Speech Technology** 2/3

Gerstendörfer, R.K. and Rohr, G. (1987)Which task representation on what type of interface In Ed. H.J.Bullinger and B. Shackel **Interact '87** Amsterdam: Elsevier 513-518

Giles, H. and Powisland, P.F. (1975) **Speech Styles and Social Evaluation** London: Academic Press

Gold, B. (1966) Word recognition copmuter program **MIT Report no.452** 1-21 Cambridge, Mass: MIT

Goodstein, L. (1981) Discrimantive display support for process operators In Ed. J. Rasmussen and W.B. Rouse **Human Detection and Diagnosis of System Failures** New York: Plenum Press

Gould, J.D. (1978) How experts dictate **Journal of Experimental Psychology: Human Perception and Performance** 4 (4) 648-661

Gould, J.D., Conti, J. and Hovanyecz, T. (1983) Composing letters with a simulated listening typewriter **Communications of the ACM** 26 (4) 295-308

Gray, J.A. and Wedderburn, A.A. (1960) Grouping strategies with simultaneous stimuli **Quarterly Journal of Experimental Psychology** 12 180-184

Green, T.R.G., Payne, S.J., Morrison, D.L., and Shaw, A. (1983) Friendly interfacing to simple speech recognisers **Behaviour and Information Technology** 2 23-38

Green, T.R., Payne, S.J., and van der Veer, G.C. (1983) **The Psychology of Computer Use** New York: Academic Press

Greenwald, A.G. (1970) A choice reaction time test of ideomotor theory **Journal of Experimental Psychology** 86 20-25

Grice, H.P. (1975) Logic and communication In Ed. P. Cole and J.L. Morgan **Syntax and Semantics: Vol.3 Speech Acts** New York: Academic Press

Grimson, W.E.L. and Patil, R.S. (1987) **A.I. in the 1980s and Beyond: an M.I.T Survey** Cambridge, Mass.: M.I.T. Press

Grossberg, S. (1987) **The Adaptive Brain, II: Vision, Speech, Language, and Motor Control** Amsterdam: North-Holland

Grossberg, S. (1988) **Neural Networks and Natural Intelligence** Cambridge, Mass: MIT Press

Grosz, B.J. (1977) **The Representation and Use of Focus in Dialogue Understanding** Menlo Park, CA: AI Centre Technical Report 151

Grosz, B.J. (1981) Focusing and description in natural language dialogues In Ed. A. Joshi et al **Elements of Discourse Understanding** Cambridge: Cambridge University Press

Grosz, B.J. and Sidner, C.L. (1986) Attention, intention and the structure of discourse **Computational Linguistics** 12 (3) 175-204

Guastello, S.J., Traut, M., and Korinenek, G. (1989) Verbal vs. pictorial representation of objects in a human computer interface **International Journal of Man Machine Studies** 31 99-120

Guenther, R.K., Klatzky, R.L. and Putnam, W. (1980) Commonalities and differences in semantic decisions about pictures and words **Journal of Verbal Learning and Verbal Behaviour** 19 54-74

Haigh, R. and Clarke, A.K. (1988) Evaluation of a voice recogntion system for use by disabled people In Ed. E.D. Megaw **Contemporary Ergonomics 1988** 87-91

Hale, A.R. and Glendon, A.I. (1987) **Individual Behaviour in the Control of Danger** Amsterdam: Elsevier

Halliday, M.A.K. (1967) Notes on transitivity and theme in English: II **Journal of Linguistics** 3 199-244

Halliday, M.A.K. (1973) **Explorations in the Functions of Language** London: Edward Arnold

Hapeshi, K., Hudson, S. and Jones, D.M. (1988) Voice data entry feedback and short term memory In Ed. E.D. Megaw **Contemporary Ergonomics 1988** London: Taylor and Francis 105-110

Hapeshi, K. and Jones, D.M. (1989) The ergonomics of automatic speech recognition interfaces **International Review of Ergonomics** 2 251-290

Hapeshi, K. and Jones, D.M. (1989b) Concurrent manual tracking and speaking: implications for automatic speech recognition In Ed. M.J. Smith and G. Salvendy **Work with Computers: Management, Stress and Health Aspects** Amsterdam: Elsevier

Hartson, H.R. (1986) **Advances in Human Computer Interaction vol. I** Norwood, NJ: Ablex

Harris, S.D., North, R.A. and Owens, J.M. (1978) A system for the assessment of human performance in concurrent verbal and manual control tasks **Behaviour Research Methods and Instrumentation** 10 (2) 329-333

Harvey, N. (1988) Are models of the future used to anticipate targets in tracking tasks? In Ed. A.M. Colley and J.R. Beech **Cognition and Action in Skilled Behaviour** Amsterdam: North Holland

Haton, J.P. (1982)**Automatic Speech Analysis and Recognition** Dordrecht: Reidel

Hauptmann, A.G. and Rudnicky, A.I. (1988) Talking to computers: an empirical investigation **International Journal ofMan Machine Studies 28** 583-604

Hayes, P.J., Hauptman, A.G. and Carbonnell, J.G. (1986) Parsing spoken language: a semantic caseframe approach **Proceedings of COLING** 587-592

Hecker, M.H., Stevens, K.N., von Bismark, G. and Willimas, C.E. (1968) Manifestations of task induced stress in the acoustical stress signal **Journal of the Acoustical Society of America 44** 993-1001

Helander, M., Moody, T.S. and Joost, M.G. (1988) System design for automated speech recognition In Ed. M. Helander **Handbook of Human Computer Interaction** Amsterdam: Elsevier

Hershman, R.L. and Hillix, W.A. (1965) Data processing in typing: typing rate as a function of kind of material and amount exposed **Human Factors 7** 483-492

Hick,W.E. (1952) On the rate of gain of information **Quartely Journal of Experimental Psychology 4** 11-26

Hickey, A.E. and Blair, W.C. (1958) Man as monitor **Human Factors 1 (1)** 8-15

Hill, D.R. (1980) Spoken language generation and understanding by machine: a problems and applications oriented overview In Ed. J.C. Simon **Spoken Language Generation and Understanding** Dordrecht: D. Reidel

Hillinger, M.L. (1980) Priming effects with phonemically similar words: the encoding bias hypothesis revisited **Memory and Cognition 8** 115-123

Hitt, W.D. (1961) An evaluation of five different abstract coding methods- experiment IV **Human Factors 3** 120-130

Hoc, J-M. (1987) Analysis of cognitive activities in process control for the design of computer aids In Ed. H-J. Bullinger and B. Shackel **Interact'87** Amsterdam: Elsevier 257-262

Hoffman, P.J., Earle, T.C., and Slovic , P. (1981) Multidimensional functional learning and some new conceptions of feedback **Organisational Behaviour and Human Performance 27** 75-102

Holding, D.H. (1965) **Principles of Training** London: Pergamon Press

Holmes, J.N. (1984) Forward to Proceedings of 1st. International Conference on Speech Technology

Holmes, J.N. (1988) **Speech Synthesis and Speech Recognition** Wokingham: van Nostrand Reinhold

Hollnagel, E., Mancini, G. and Woods, D.D. (1988) **Cognitive Engineering in Complex Dynamic Worlds** London: Academic Press

Hopkins, R.H. and Atkinson, R.C. (1968) Priming and the retrieval of names from long term memory **Psychonomic Science 11** 219-220

House, A.S., Williams, C.E., Hecker, M.H.L, and Kryter, K. (1965) Articulation testing methods: consonant differentiation with a closed response set **Journal of the Acoustical Society of America 37** 158-166

Hubel, D.H. and Weisel, T.N. (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex **Journal of Physiology 33A** 106-154

Itakura, F. (1975) Minimum prediction residual principle applied to speech recognition **IEEE Transactions ASSP 23** 67-72

Ito, N., Inoue, S., Ohkura, M. and Masada, W. (1989) The effect of voice message length on the interactive copmuter systems In Ed. F. Klix et al. **Man Computer Interaction Research - Macinter II** Amsterdam: North Holland 245-252

Jack, M.A. and Laver, J. (1988) **Aspects of Speech Technology** Edinburgh: Edinburgh University Press

Jacobs, R.J., Johnston, A.W. and Cole, B.C. (1975) The visibility of alphabetic and symbolic traffic signs **Australian Road Research** 5 (7) 68-86

Jaffe, J. and Feldstein, S. (1970) **Rhythms of Dialogue** London: Academic Press

Jakobson, R., Fant, G., and Halle, M. (1963) **Preliminaries to Speech Analysis** Cambridge, Mass: MIT Press

Jelinek, F. (1976) Continuous speech recognition by statistical methods **Proceedings of the IEEE** 64 (4) 532-556

Jensen, R.S. (1989) **Aviation Psychology** Aldershot: Gower Technical

Johnson-Laird, P.N. (1983) **Mental Models** Cambridge: Cambridge University Press

Johnston, R.D. (1986) Speech I/O: the users requirement **Voice Processing** Pinner: Online

Johnston, W.A. and Wilson, J. (1980) Perceptual processing of non targets in an attention task **Memory and Cognition** 8 372-377

Jones, D.M., Hapeshi, K. and Frankish, C. (1987) Human factors and the problems of evaluation in the design of speech systems interfaces In Ed. D. Diaper and R. Winder **People and Computers III** Cambridge: Cambridge University Press

Jones, D.M., Hapeshi, K. and Frankish, C. (1989) Design guidelines for speech recognition interfaces **Applied Ergonomics 20** 47-52

Joshi, A., Webber, B., and Sag, I. (1981) **Elements of Discourse Understanding** Cambridge: Cambridge University Press

Kahneman, D. (1973) **Attention and Effort** Englewood Cliffs, NJ: Prentice Hall

Karhan, C.J. (1987) Human factors issues in applying ASR to network services **10th.International Symposium on Human Factors in Telecommunications**

Kawabata, T., Makino, S. and Kido, K. (1984) Four layer model of consonant recognition using phonetic features **Transactions of the Institute for Electrical and Communications Engineering in Japan** j67d (1) 141-148

Keele, S.W. (1973) **Attention and Human Performance** Englewood Cliffs, N.J.: Prentice Hall

Kelley, C. (1968) **Manual and Automatic Control** New York : Wiley

Kelley, J.F. and Chapanis, A. (1977) Limited vocabulary natural language dialog **International Journal of Man Machine Studies** 9 479-501

Kelley, J.F. (1983) An empirical methodology for writing user friendly natural language computer applications **Proceedings of CHI'83 Conference on Human Factors in Computing Systems**

Kelway, P. (1988) Speech technology - a slow revolution **Chartered Mechanical Engineering**

Kersteen, Z.A. and Damos, D. (1983) **Human Factors Issues Associated with the Use of Speech Technology in the Cockpit** Arizona State University Report No.: NASA CR 166548

Kidd, A.L. (1982) Problems of man machine dialogue design **Proceedings 6th. International Conference on Computer Communication** 531-536

Kinchla, R.W. (1980) The measurement of attention In Ed. R. Nickerson **Attention and Performance VIII** Hillsdale, NJ: Erlbaum

Klapp, S.T. and Netick, A. (1988) Multiple resources for processing and storage in short term working memory **Human Factors 30** pp.617-632

Klatt, D.H. (1977) Review of the ARPA speech understanding project **Journal of the Acoustical Society of America 62 (6)** 1345

Klatt, D.H. (1980) Overview of the ARPA speech understanding project In Ed. W.A. Lea **Trends in Speech Recognition** New York: Prentice Hall

Klatt, D.H. and Stevens, K.N. (1973) On the automatic recognition of continuous speech: implications from a spectrogram reading experiment **IEEE Transactions Audio and Electroacoustics AU - 21** 210-217

Knapp, M.L. and Miller, G.R. (1985) **Handbook of Interpersonal Communication** Beverly Hills, CA: Sage

Knight, J. A. and Peckham, J.B. (1984) **A Generic Model for the Assessment of Speech Input Applications** Cambridge: Logica Report

Koenig, W., Dunn, H.K. and Lacey, L.Y. (1946) The sound spectrograph **Journal of the Acoustical Society of America 18** 19-49

Kohda, M.S. and Saito, S. (1973) Influence of long term variations of learning and unknown samples on recognition rate of spoken digits **Report from Autumn Meeting of Acoustical Society of Japan** 141-142

Kolers, P.A., Wrolstead, M. and Bouma, H. (1979) **Processing of Visible Language vol I** New York: Plenum Press

Kornblum, S.(1973) **Attention and Performance IV** New York: Academic Press

Kornblum, S., Hasbroucq, T. and Osman, A. (1990) Dimensional overlap: cognitive basis for stimulus response compatibility - a model and taxonomy **Psychological Review 97 (2)** 253-270

Kosslyn, S.M. and Chabris, C.F. (1990) Naming pictures **Journal of Visual Languages and Computing 1** 77-95

Kraat, A. (1985) Communication interaction between aided and natural speakers **Report from International Project on Communication Aids for the Severely Impaired**

Kragt, H. and Landeweerd, J.A. (1974) Mental skills in process control In Ed. E. Edwards and F.P.Lees **The Human Operator in Process Control** London: Taylor and Francis

Krauss, R.M. and Glucksberg, S. (1974) Social and nonsocial speech **Scientific American** 100-105

Kryter, K.D. (1972) Speech communication In Ed. H.P. van Cott and R.G. Kinkade **Human Engineering Guide to Equipment Design** New York : McGraw Hill

Kurath, H. (1949) **Word Geography of the Eastern United States** Ann Arbor, MI: University of Michigna Press

Kuroda, I., Fujiwara, O., Okamura, N. and Utsuku, N. (1976) Method for determining pilot stress through analysis of voice communication **Aviation, Space and Environmental Medicine** May 528-533

Landeweerd, J.A. (1979) Internal representation of a process, fault diagnosis and fualt correction **Ergonomics** 22 (12) 1343-1351

Lawrence, D.H. (1971).Two studies of visual search for word targets with controlled rates of presentation **Perception and Psychophysics** 10 85-89

Lea, W.A. (1980a) The value of speech recognition systems In W.A. Lea Ed. **Trends in Speech Recognition** New York: Prentice-Hall

Lea W.A. (1980b)**Trends in Speech Recognition** New York: Prentice-Hall

Lea, W.A. (1982a) Selecting, designing and using practical speech recognisers In Ed. J.P. Haton **Automatic Speech Analysis and Recognition** Dordrecht: Reidel

Lea, W.A. (1982b) Problems in predicting performance of speech recognizers In D.S. Pallett Ed. **Proceedings Workshop on the Standardization of Speech I/O Technology** Gaitherzburg, M.D.: National Buruea of Standards 15- 24

Lea, W.A. (1986) The elements of speech recognition In Ed. G. Bristow **Electronic Speech Recognition** London: Collins

Lea, W.A. and Woodard, J. (1983) New procedures for comprehensive assessment of voice entry systems **Proceedings 1983 Voice Data Entry Systems Applications Conference**

Lee, B.S. (1950) Effects of delayed speech feedback **Journal of the Acoustical Society of America** 22 824-826

Lermon, L. (1980) Traceability via voice data entry **IEEE Transactions CHMT 3** 360-366

Lees, D. S., Crigler, B., van der Loos, M., and Leifer, L. (1988) Design of control and user interface software for a third generation desktop robotic assistant In Ed. G. Curtis **Rehabilitation Research and Development Centre Progress Report** Palo Alto, CA: Veterans Administration Medical Centre

Leiser, R.G. (1989a) Improving natural language speech interfaces by the use of metalinguistic phenomena **Applied Ergonomics** 20 (3) 168-173

Leiser, R.G. (1989b) Exploiting convergence to improve natural language understanding **Interacting with Computers** 1 (3) 284-298

Leiser, R.G., de Alberti, M. and Carr, D.J. (1987) Generic issues in dialogue design for speech input/output In Ed. J. Laver and M.A. Jack **Aspects of Speech Technology** Edinburgh: CEP Consultants

Lettvin, J.Y., Maturana, H.R., McCulloch, W.S. and Pitts, W.H. (1959) What the frog's eye tells the frog's brain **Proceedings of the Institute of Radio Engineers** 47 140-151

Levin, H. and Lord, W. (1975) Speech pitch as an emotional state indicator **IEEE Transactions SMC - 5** (2) 259-272

Levinson, S.E. and Liberman, M.Y. (1981) Speech recognition by computer **Scientific American** 244

Lewis, C. and Norman, D.A. (1986) Designing for error In Ed. D.A. Norman and S.W. Draper **User Centred Design** Hillsdale, NJ: Erlbaum

Liberman, A.M., Cooper, F.S., Shankweiler, D.P.and Studdert Kennedy, M. (1967) Perception of the speech code **Psychological Review 74** 431-461

Lierberman, P. and Michaels, S.B. (1962) Some aspects of fundamental frequency and envelope amplitue related to the emotional content of speech **Journal of the Acoustical Society of America 34** (7) 922-927

Lieberman, P. and Blumstein, S. (1988) **Speech Physiology, Speech Perception and Acoustic Phonetics** Cambridge: Cambridge University Press

Linde, C. and Shively, R.J. (1988) Field study of communication and workload in police helicopters: implications for cockpit design **Proceedings Human Factors Society 32rd.Annual Meeting** Santa Monica, CA: Human Factors Society

Lindgren, N. (1965) Machine recognition of human language **IEEE Spectrum vol. 2** (March, April, May)

Linggard, R. (1990) Beyond speech recognition: language processing In Ed. C.Wheddon and R.Linggard **Speech and Language Processing** London: Chapman and Hall

Little, A. and Joost, M.G. (1984) A comparison of two error correction stratgies in voice data entry **Proceedings AVIOS '84: Voice I/O Systems Applications Conference** Arlington, VA

Little, R. and Cowan, R. (1986) **A Flight Evaluation of Voice Interaction as a Component of an Integrated Helicopter Avionics System** Farnborough: Royal Aircraft Establishment Technical Memorandum FS(B) 637

Loftus, G.R. and Bell, S.M. (1975) Two types of information in picture memory **Journal of Experimental Psychology: Human Learning and Memory 1** 103-115

Longacre, R.E. (1983) **The Grammar of Discourse** New York: Plenum Press

Lowerre, B.T. (1976) **The HARPY speech recognition system** Unpublished PhD Thesis Carnegie Mellon Univesity

Luce, P.A., Feustel, T.C., and Pisoni, D.B. (1983) Capacity demands in short term memory for synthetic and natural speech **Human Factors 25** 17-32

Lupker, S.J. and Williams, B.A. (1989) Rhyme priming of picture and words: a lexical activation account **Journal of Experimental Psychology:Learning, Memory and Cognition 15** (6) 1033-1046

Makhoul, J. (1975) Linear prediction of speech: a tutorial review **Proceedings of the IEEE 63** 561-580

Malhotra, A. (1975) **Design Criteria for a Knowledge based English Language System for Management:an Experimental Analysis** Cambridge, Mass: MIT Report TR - 146

Mann, T.L. and Hammer, J.M. (1986) Analysis of user procedural compliance in controlling a simulated process **IEEE Transaction SMC - 16** (4) 505-510

Markel, J.D. and Gray, A.H. (1976) **Linear Prediction of Speech** Berlin: Springer Verlag

Marks, W. (1989) Elaborative prcoesing of pictures in verbal domains **Memory and Cognition 17** (6) 662-672

Marshall, E. and Shepherd, A. (1977) Strategies adopted by operators when diagnosing plant failures from a simulatedcontrol panel In **Human Operators and Simulation** London: Institute of Measurement and Control

Marslen Wilson, W.D. (1987) Functional parallelism in spoken word recognition In Ed. U.H. Frauenfelder and L.K. Tyler **Spoken Word Recognition** Cambridge, Mass: MIT Press

Marslen-Wilson, W.D. and Welch, A. (1978) Proceesing interaction and lexical access during word recognition in continuous speech **Cognitive Psychology 10** 29- 63

Martin, G.L. (1989) The utility of speech input in user - computer interfaces **International Journal of Man Machine Studies 30** 355-375

Martin, T.B. (1976) Practical applications of voice input to machines **Proceedings IEEE 64** (4) 487-501

Martin, T.B. and Welch, J.R. (1980) Practical speech recognisers and some performance effectiveness parameters In W.A. Lea Ed. **Trends in Speech Recognition** Englewood Cliffs, N.J.: Prentice-Hall

Masson, M.E.J. (1983) Conceptual processing of text during skimming and rapid sequential reading **Memory and Cognition 11** 262-274

Matthei, E. and Roeper, T. (1983) **Understanding and Producing Speech** London: Fontana

McCandless, S. S. (1974) An algorithm for automatic formant extraction using linear prediction spectra **IEEE Transactions ASSP** 22 (2) 135-141

McCauley, M.E. (1984) Human factors in voice technology In Ed. F.A. Muckler **Human Factors Review: 1984** Santa Monica,CA: Human Factors Society

McCauley, M.E. and Semple, C.A. (1980) **Precision Approach Radar Training System** Orlando, Fl: Naval Training Equipment Centre Report NAVTRAEQUIPCEN 79 - C -0042 -1

McInnes, F.R. and Jack, M.A. (1987) Reference template adaptation in speaker indepedent isolated word speech recognition **Electronics Letters 23** (24) 1304-1305

McInnes, F. and Jack, M.A. (1988) ASR using word reference patterns In Ed. M.A. Jack and J. Laver **Aspects of Speech Technology** Edinburgh: Edinburgh University Press

McMahon, M.L. (1984) Talking to your PC or what do you do after it says 'Hello'? **Proceedings SpeechTech'84** 127-133

Mead, M. and Modley, R. (1968) Communications among all people,everywhere **Natural History 77** 56-63

Meisel, W.S. (1986) Towards the 'talkwriter' In G. Bristow **Electronic Speech Recognition** London: Collins

Metton, A.W. (1964) **Categories of Human Learning** New York: Academic Press

Meyer, D.E. (1981) Latency differences in monoptic and dichoptic shape and colour decision making **Journal of Experimental Psychology: Human Perception and Performance 7** 968-971

Meyer, D.E. and Schvanveldt, R.W. (1971) Facilitation in recognising pairs of words: evidence of a dependence between retrieval operations **Journal of Experimental Psychology 90** 227-235

Meyer, D.E., Schvaneveldt, R.W. and Ruddy, M.G. (1974) Functions of graphemic and phonemic codes in visual word recognition **Memory and Cognition** 2 309-321

Miller, G. A., Heise, G., and Lichten, W. (1951) The intelligibility of speech as a function of the context of speech materials **Journal of Experimental Psychology** 41 329-335

Miller, G.A. and Nicely, P.E. (1955) An analysis of perceptual confusions among some English consonants **Journal of the Acoustical Society of America** 27 338-352

Mitchell, C.M. and Forren, M.G. (1987) Mulitimodal user input to supervisory control systems: voice augmented keyboard **IEEE Transactions SMC** 17 (4) 594-607

Monk, A. (1984) **Fundamentals of Human Computer Interaction** London: Academic Press

Moore, R.K. (1977) Evaluating speech recognisers **IEEE Transactions ASSP** 25 178-183

Moore, R.K. (1984) Overview of speech input In Ed. J.N. Holmes **Proceedings of the 1st. International Conference on Speech Technology** Amsterdam: North Holland

Moore, T.J. (1989) Speech technology in the cockpit In Ed. R.S. Jensen **Aviation Psychology** Aldershot: Gower Technical

Moore, T.J. and McKinley, R.L. (1986) Research on speech processing for military avionics **Proceedings Human Factors Society 30th. Annual Meeting** 1331-1335

Moray, N. (1959) Attention in dichotic listening: affective cues and the influence of instructions **Quarterly Journal of Experimental Psychology** 11 56-60

Moray, N. (1976) Attention, control and sampling behaviour In Ed. T.B. Sherian and G.Johannsen **Montoring Behaviour and Supervisory Control** New York: Plenum Press

Moray, N. (1981a) The role of attention in the detection of errors and the diagnosis of failures in man machine systems In Ed. J. Rasmussen and W.B. Rouse **Human Detection and Diagnosis of System Failures** New York: Plenum Press

Moray, N. (1981b)Feedback and the control of skilled behaviour In Ed. D.H. Holding **Human Skills** Chichester: Wiley

Moray, N. and Rotenberg, I. (1989) Fault management in process control: eye movement and action **Ergonomics** 32 (11) 1319-1342

Morris, N. and Jones, D.M. (1987) Reporting words from the eye or the ear: to write or to speak? **Ergonomics** 30 665-674

Morris, N.M. and Rouse, W.B. (1985) The effects of type of knowledge upon human problem solving in a process control task **IEEE Transaction SMC** - 15 (6) 698-707

Morton, J. (1969) Interaction of information in word recognition **Psychological Review** 76 165-178

Morton, J. (1979) Facilitation in word recognition: experiments causing change in the logogen model In Ed. P.A. Kolers et al **Processing of Visible Language** vol. I New York : Plenum Press

Mulla, H. (1984) An experimental voice command system for PABX and automated office application **Proceedings SpeechTech'84** 48-52

Müller, J. (1848) **The Physiology of the Senses, Voice, and Muscular Motion with the Mental Faculties** London: Walton and Maberly

Mullins, P.A. (1988) Considerations of intention and movement: action plans and motor programs in speech and keyboard use **Proceedings Human Factors Society 32nd. Annual Meeting** 549-553

Murrell, H. (1976) **Men and Machines** London: Methuen

Mutschler, H. (1982) Ergonomic aspects for improving recognition performance of voice input systems **IFAC Analysis, Design and Evaluation of Man Machine Systems** 261-267

National Research Centre (1983) **Research Needs for Human Factors** Washington, DC: National Academy Press

Navon, D. and Gopher, D. (1979) On the economy of the human processing system **Psychological Review** 86 (3) 214-255

Neisser, U. (1963) Decision time without reaction time: experiments in visual scanning **American Journal of Psychology** 76 376-385

Nelson, D.L., Reed, V.S. and McEvoy, C.L. (1977) Learning to order pictures and words: a model of sensory and semantic encoding **Journal of Experimental Psychology: Human Learning and Memory** 3 485-497

Newell, A. (1975) A tutorial on speech understanding systems In Ed. D. Raj Reddy **Speech Recognition: Invited Papers Presented at the 1974 Symposium** New York: Academic Press

Newell, A., Barnett, J., Forgie, J.W., Green, C., Klatt, C., Licklider, J.C.R., Munson, J., Reddy, D.R. and Woods, W.A. (1973) **Speech Understanding Systems: Final Report of a Study Group** New York: North Holland

Newell, A.F. (1984) Speech: the natural method of man machine communication **Proceedings 1st. IFIP Conference on Human Computer Interaction** 231-238

Newell, A.F. (1986) Communicating via speech - the able bodied and disabled **IEE Conference no. 258** London: IEE

Newell, A.F., Arnott, J.L, and Dye, R. (1987) A full speed simulation of speech recognition machines **European Conference on Speech Technology** 1-4

Nickerson, R.W. (1978) On the time it takes to tell things apart In Ed. M. Gazzaniga **Handbook of Behavioural Neurobiology II Neuropsychology** New York: Plenum Press

Nolan, F. (1983) **The Phonetic Bases of Speaker Recognition** Cambridge: Cambridge University Press

Nolan, F. (1986) The nature of speech In Ed. G. Bristow **Electronic Speech Recognition** London: Collins

Norman, D.A. (1968) Toward a theory of memory and attention **Psychological Review** 75 522-536

Norman, D.A. (1988) **The Psychology of Everyday Things** New York : Basic Books

Norman, D.A. and Rumelhart, D.E. (1975) **Explorations in Cognition** San Francisco: Freeman

North, R. and Lea, W.A. (1982) **Application of Advanced Speech Technology in Manned Penetration Bombers** Ohio: Wright Patterson Aeronautical Laboritories Report No.: AFWAL TR 82 3004

Noyes, J.M. and Frankish, C.F. (1986) Voice recognition - where are the end users? **Proceedings European Conference on Speech Technology** London: CEP Consultants 349-352

Noyes, J.M. and Frankish, C. R. (1989) A review of speech recognition applications in the office **Behaviour and Information Technology 8** (6) 475-486

Noyes, J. M., Haigh, R. and Starr, A.F. (1989) Automatic speech recognition for disabled people **Applied Ergonomics 20** (4) 293-298

Nye, J.M. (1980) Expanding markets for ASR In Ed. W.A. Lea **Trends in Speech Recognition** Hillsdale, NJ: Prentice Hall

Nye, J.M. (1982) Human factors analysis of speech recognition systems **Speech Technology** (April) 50-57

Ohala, J.J. (1982) Calibrated vocabularies In Ed. D.S. Pallett **Proceedings Workshop on the Standardization of Speech I/O Technology** Gaitherzburg, M.D.: National Buruea of Standards

Oldfield, R.C. and Wingfield, A. (1964) The time it takes to name an object **Nature 202** 1031-1032

Olson, D.R. (1970) Language and thought: aspects of a cognitive theory of semantics **Psychological Review 77** 257-273

Olson, H. and Belar, H. (1956) Phonetic typewriter **Journal of the Acoustical society of America 28** 1072-1081

Paivio, A. (1986) **Mental Representation: A Dual Coding Approach** Oxford: Oxford University Press

Pallett, D.S. (1982) **Proceedings Workshop on the Standardization of Speech I/O Technology** Gaitherzburg, M.D.: National Buruea of Standards

Parsons, T.W. (1987) **Voice and Speech Processing** New York: Mc.Graw Hill

Paternotte, P.H. (1978) The control performance of operators controlling a continuous distillation process **Ergonomics 21** 671-

Peckham, J.B. (1984) Speech recognition - what is it worth? In Ed. J.N. Holmes **Proceedings of the 1st. International Conference on Speech Technology** Amsterdam: North Holland

Peckham, J.B. (1985) Speech technology assessment activities in the UK **SpeechTech '85** 165-169

Peckham, J. (1986) Human factors in speech recognition In Ed. G. Bristow **Electronic Speech Recognition** London: Collins

Peckham, J. (1989) **Recent Developments and Applications of Natural Language Processing** London: Kogan Page

Peckham, J.B. and Knight, J.A. (1984) **A Generic Model for the Assessment of Speech Input Applications** Cambridge: Logica Report

Pick, H.L. and Acredelo, L.P. (1983) **Spatial Orientation: Theory, Research and Application** New York: Plenum Press

Pickett, J.M. (1980) **The Speech Sounds of Commuincation. A Primer of Acoustic Phonetics and Speech Perception** Baltimore: University Park Press

Pinsky, L. (1983) What kind of "dialogue" is it when working with a computer In Ed. Green et al **The Psychology of Computer Use** London: Academic Press

Pisoni, D.B. (1982) Perception of speech: the human listener as a cognitive interface **Speech Technology** (April) 10-23

Pols, L.C.W. and Plomp, R. (1986) How to make more efficient use of the fact that the speech signal is dynamic and redundant **Proceedings International Conference ASSP 86** 1963-1965

Poock, G.K. (1980) **Experiments With Voice Input for Command and Control** Monterey: Naval Postgraduate School Tech. Rep. NPS-55-80-016

Poock, G.K. (1981) **A Longitudinal Study of Computer Voice Recognition Performance and Vocabulary Size** Monterey: Naval Postgraduate School Tech. Rep. NPS-55-81-013

Poock, G.K. (1982) Using voice input to operate a distributed computer network **Proceedings of Conference on Voice - Interactive Systems: Applications and Payoffs** 213-229

Poock, G.K., Martin, B.J. and Roland, E.F. (1983) **The Effect of Feedback to Users of Voice Recognition Equipment** Monterey: Naval Postgraduate School Tech. Rep. NPS-55-31-003

Posner, M.I. and Snyder, C.R. (1975) Attention and cognitive control In Ed. R. Solso **Information Processing and Cognition: The Loyola Symposium** Hillsdale, NJ: Lawrence Erlbaum

Potter, M.C. and Faulconer, B.A. (1975) Time to understand pictures and words **Nature 253** 437-438

Potter, J. and Wetherell, M. (1987) **Discourse and Social Psychology** London: Sage

Poulton, E.C. (1982) Influential companions: effects of one strategy on another in the within subjects designs of cognitive psychology **Psychological Bulletin 91** 673-690

Poulton, A.S. (1983) **Microcomputer Speech Synthesis and Recognition** Chichester: John Wiley

Price, J. R. (1969) Whither speech recognition? **Journal of the Acoustical Society of America 46** (2)

Pylyshyn, Z. (1986) **Computing and Cognition** Cambridge, Mass: MIT

Quarmby, D. (1986) Silicon devices for speech recognition In Ed. G. Bristow **Electronic Speech Recgonition** London: Collins

Rasmussen, J. (1974) On the communication between operators and instrumentation in automatic process plants In Ed. E.Edwards and F.P. Lees **The Human Operator in Process Control** London: Taylor and Francis

Rasmussen, J. (1976) Outlines of a hybrid model of the process plant operator In Ed. T.B. Sheridan and G. Johannsen **Monitoring Behaviour and Supervisory Control** New York: Plenum Press

Rasmussen, J. (1981) Models of mental strategies in process plant diagnosis In Ed. J. Rasmussen and W.B. Rouse **Human Detection and Diagnosis of System Failures** New York : Plenum Press

Rasmussen, J. (1983) Skills, rules, and knowledge; signals, signs and symbols, and other distinctions in human performance models **IEEE Trans. Systems, Man, and Cybernetics SMC-13** (3) pp.257-266

Rasmussen, J., Duncan, K., and Leplat, J. (1987) **New Technology and Human Error** Chichester: John Wiley

Rasmussen, J. and Jensen, A. (1974) Mental procedures in real life tasks: a case study of electronic trouble shooting **Ergonomics 17** (3) 293-307

Rasmussen, J. and Rouse, W.B. (1981) **Human Detection and Diagnosis of System Failures** New York : Plenum Press

Ratcliff, R. (1985) Theoretical interpretations of the speed and accuracy of positive and negative responses **Psychological Review 92** 212-225

Reason, J.T. (1979) Actions not as planned In Ed. G. Underwood and R. Stevens **Aspects of Consciousness** London: Academic Press

Reason, J. T. and Mycielska, K. (1982) **Absent Minded? The Psychology of Mental Lapses and Everyday Errors** New Jersey: Prentice-Hall

Reddy, D.R. (1976) Speech recognition by machine: a review **Proceedings of the IEEE 64** 501-531

Reddy, D.R. (1980) Models of speech perception In Ed. R.A. Cole **Perception and Production of Fluent Speech** Hillsdale, NJ: Erlbaum

Reed, L. (1985) Military applications of voice technology **Speech Technology** 2 (4) 42-50

Rehsöft, C. (1984) Voice recognition at the Ford warehouse in Cologne In Ed. J.N. Holmes **Proceedings of the 1st. International Conference on Speech Technology** Amsterdam: North Holland 103-112

Reilly, R.G. (1987) Types of communication failure in dialogue In Ed. R.G. Reilly **Communication Failure in Dialogue and Discourse: Detection and Repair Processes** Amsterdam: North Holland

Reilly, R.G. (1987) **Communication Failure in Dialogue and Discourse: Detection and Repair Processes** Amsterdam: North Holland

Reising, I.M, and Curry, D.M. (1987) Comparison of voice with multifunction controls: logic is the key **Ergonomics 30** 1063-1078

Reisner, P. (1977) Use of psychological experimentationas an aid to development of a query language **IEEE Transactions** SE 3 218-220

Remington, R. and Williams, D. (1986) On the selection and evaluation of visual display symbology : factors influencing search and identification times **Human Factors 28** 407-420

Richards, M.A. and Underwood, K.M. (1984) Talking to machines: How are people naturally inclined to speak? In Ed. E.D. Megaw **Contemporary Ergonomics 1984.** (London : Taylor and Francis)

Richardson Simon, J., Peterson, K.D., and Wang, J.H. (1988) Same-different reaction times to stimuli presented simultaneously to separate cerebral hemispheres **Ergonomics 31** 1837-1846

Ringle, M.D. and Bruce, B. (1982) Conversation failure In Ed. W. Lehnert and M. D. Ringle **Strategies for Natural Language Processing** Hillsdale, NJ: Erlbaum

Ringle, M.D. and Halstead Nussloch, R. (1989) Shaping user input: a strategy for natural language dialogue design **Interacting with Computers** 1 (3) 227-244

Robinson, C.R. and Eberts, R.E. (1987) Comparison of speech and pictorial displays in a cockpit environment **Human Factors 29** 31-44

Rollins, A.M. (1984) Speech recognition and manner of speaking in noise and quiet **Proceedings CHI' 85** New York: Association for Computing Machinery

Rosinski, R.R., Chiesi, H. and Debons, A. (1980) Effects of amount of visual feedback on typing performance **Proceedings Human Factors Society 24th. Annual Meeting** 195-199

Rotter, J.B. (1966) Generalized expectencies for internal vs.external control of reinforcement **Psychological Monographs 609**

Rubenstein, H., Garfield, L. and Millikan, J.A. (1970) Homographic entries in the internal lexicon **Journal of Verbal Learning and Verbal Behaviour** 9 487-492

Rutter, D.R. (1987) **Communicating by Telephone** Oxford : Pergamon Press

Sacks, H., Schlegoff, E. and Jefferson, G. (1978) A simplest systematics for the organization of turn-taking for conversation In Ed. J. Schenkein **Studies in the Organization of Conversational Interaction** New York : Academic Press

Salvendy, G. (1987) **Handbook of Human Factors** New York: John Wiley

Schank, R. C. and Abelson, R. P. (1977) **Scripts, Plands, Goals and Understanding** Hillsdale, NJ: Erlbaum

Schenkein, J. (1980) A taxonomy of repeating action sequences in natural conversation In Ed. B. Butterworth **Language Production I Speech** New York: Academic Press

Scherer, K.R. (1981) Vocal indicators of stress In Ed. J. Darby **Speech Evaluation in Psychiatry** New York: Grune and Stratton

Scherer, K.R. (1986) Vocal affect expression: a review anda model for future research **Psychological Review** 99 (2) 143-165

Schmadt, C. (1986) Voice communication with computers In Ed. H.R. Hartson **Advances in Human Computer Interaction I** Norwood, NJ: Ablex

Schurick, J.M. (1986) Efficiency of limited vocabulary speech recognition for data entry tasks **Proceedings Human Factors Society 30th. Annual Meeting** 931-934

Schurick, J.M., Williges, B.H. and Maynard, J.F. (1985) User feedback requirements with ASR **Ergonomics** 28 1534-1555

Schwab, E.C., Nusbaum, H.C., and Pisoni, D.B. (1985) Effects of training on the perception of synthetic speech **Human Factors** 27 (4) 395-408

Searle, J. R. (1969) **Speech Acts** Cambridge: Cambridge University Press

Sedgwick, N. (1987) Speech recognition technology: how well does it satisfy the need? **Proceedings International SpeechTech '87** New York: Media Dimensions

Selfridge, O.G. (1959) Pandemonium: a paradigm for learning In **Mechanisation of Thought Processes** ·London: HMSO

Seibel, R. (1972) Data entry devices and procedures In Ed. H.P. van Cott and R.G. Kinkade **Human Engineering Guide to Equipment Design** Washington, DC: U.S. Government Printing Office

Seidenberg, M.S. and McClelland, J.L. (1989) A distributed model of word recognition and naming **Psychological Review** 96 523-568

Shepherd, A. (1989) Analysis and training in information technology tasks In Ed. D. Daiper **Task Analysis for Human Computer Interaction** Chichester: Ellis Horwood

Shepherd, A., Marshall, E.C., Turner, A. and Duncan, K.D. (1977) Diagnosing plant failures from a control panel: a comparison of three training methods **Ergonomics** 20 347-361

Sheridan, T.B. (1988) Task allocation and supervisory control In Ed. M. Helander **Handbook of Human Computer Interaction** Amsterdam: Elsevier

Sheridan,T.B. and Johannsen, G. (1976) **Montoring Behaviour and Supervisory Control** New York: Plenum Press

Shiffrin, R.M. and Schneider, W. (1977) Controlled and automatic human information processing: II. perceptual learning, automatic attending, and a general theory **Psychological Review 84** 127-190

Shneiderman, B. (1987) **Designing the User Interface: Strategies for Human Computer Interaction** Reading, Mass: Addison Wesley

Simes, D.K. and Sirsky, P.A. (1987) Human factors: an exploration of the psychology of human computer dialogues In Ed. H.R. Hartson **Advances in Human Computer Interaction 1** Norwood, NJ: Ablex

Simonov, P.V. and Frolov, M.V. (1973) Utilisation of human voice for estimation man's emotional stress and state of attention **Aerospace Medicine** March 256-258

Simpson, C.A. (1986) Speech variability on recognition accuracy associated with concurrent task performance by pilots **Ergonomics** 29 (11) 1343-1357

Simpson, C.A., Mc.Cauley, M.E., Roland, E.F., Ruth, J.C., and Williges, B.H. (1985) System Design for Speech Recognition and Generation **Human Factors** 27 (2) 115-143

Small, D. and Weldon, L. (1983) An experiemental study of natural and structured query languages **Human Factors 25** 253-263

Smith, A.R. and Sambur, M.R. (1980) Hypothesising and verifying words for speech recognition In Ed. W.A. Lea **Trends in Speech Recognition** Englewood Cliffs, NJ: Prentice Hall

Smith, K.V. and Smith, H.M. (1962) **Perception and Analysis of Space Structured Behaviour** Philadelphia, PA: Saunders

Smith, K.V. and Smith, M.F. (1966) **Cybernetic Principles of Learning and Educational Design** New York: Holt Rineholt and Winston

Smith, S.L. and Mosier, J.N. (1984) **Design Guidelines for User System Interface Software** Maynard, Mass: Digital Corp.

Smith, T.J. and Smith, K.U. (1987) Feedback control mechanisms of human behaviour In Ed. G. Salvendy **Handbook of Human Factors** New York: John Wiley

Solzenhitsyn, A. (1968) **The First Circle** London: Penguin

Sondheimer, N.K. (1976) Spatial reference and natural language machine control **International Journal of Man Machine Studies 8** 329-336

Sperber, D. and Wilson, D. (1985) **Relevance: Communication and Cognition** Oxford: Basil Blackwell

Spine, T.M., Maynard, J.F., and Williges, B.H. (1983) Error Correction Strategies for Voice Recognition **Proceedings Voice Data Entry Systems Applications Conference** Chicago, Il: American Voice I/O Society

Spine, T.M., Williges, B.H. and Maynard, J.F. (1984) An economical approach to modelling speech recognition accuracy **International Journal of Man Machine Studies 21** 191-202

Stammers, R.B., George, D.A. and Carey, M.S. (1989) An evaluation of abstract and concrete icons for a CAD package In Ed. E.D. Megaw **Contemporary Ergonomics 1989** London: Taylor and Francis 416-421

Stanton, N. and Booth, R.T. (1990) The psychology of alarms In Ed. R.J. Lovesey **Contemporary Ergonomics 1990** London: Taylor and Francis 378-383

Starr, A.F., Hudson, S.M., and Jones, D.M. (1988) User **preferences and researchers experiences: applications of speech recognition in the United Kingdom and North America** Report 3: Alvey Contract MMI 103 Human Factors in the Design of Speech System Interfaces Cardiff: UWIST

Steiner, B.A. and Camacho, M.J. (1989) Situation awareness: icons vs. alphanumerics **Proceedings Human Factors Society 33rd. Annual Meeting** Santa Monica, CA: Human Factors Society 28-32

Stephens, R.M., Cottle, M.J., Creasey, G.H., Geggie, C.S., and Workman, D.S. (1988) Text composition using speech recognition and other computer input devices for people with spinal cord injuries **Proceedings Speech'88: 7th. FASE Symposium** 337-344

Sterling, M.J.H. (1978) **Power System Control.** Stevenage :IEE/Peter Peregrinus Ltd.

Stern, P., Eskenazi, M. and Memmi, D. (1986) An expert system for spectrogram reading **Proceedings International Conference ASSP '86** 1193-1196

Sternberg, S. (1975) Memory scanning: new findings and controversies **Quarterly Journal of Experimental Psychology** 27 1-32

Sternberg, S., Monsell, S., Knoll, R.L. and Wright, C.E. (1978) The latency and duration of rapid movement sequences: comparisons of speech and typewriting In Ed. G.E. Stelmach **Information Processing in Motor Control and Learning** New York: Academic Press

Sternberg, S., Wright, C.E., Knoll, R.L. and Monsell, S. (1980) Motor programs in rapid speech: additional evidence In Ed. R.A. Cole **The Perception and Production of Fluent Speech** Hillsdale, NJ: Erlbaum

Suchman, L. (1987) **Plans ans Situated Actions: the Problems of Human Machine Communication** Cambridge: Cambridge University Press

Szlichcinski, K.P. (1977) Symbols and pictograms: a review of their usefulness and the methodology of their design **Proceedings 8th. International Symposium on Human Factors in Telecommunications** 357-378

Talbot, M. (1986) Adapting ASR to human diction **Adaptive Man Machine Interfaces** London: IEE Colloquium 1986/110

Talbot, M. (1987) Human Speech Production and Automatic Speech Recognition: Resolving Some Differences In Ed. J.A. Waterworth **Speech and Language Based Interaction with Machines: Towards the Conversational Computer** Chichester: Ellis Horwood

Tattersall, G.D., Linford, P.W. and Linggard, R. (1988) Neural arrays for speech recognition **British Telecom Technology Journal** 6 (2) 140-163

Taylor, M.R. (1986) Direct voice input and its role in avionics systems **International Conference on Speech Technology** Oct. 1986 Brighton 113-120

Teja, E.R. and Gonnella, G.W. (1983) **Voice Technology** Reston,VA: Reston Publishing

Thomas, J.C. (1987) Tools and Methodologies for Speech Recognition Assessement **Proceedings SpeechTech'87** New York: Media Dimensions 250- 253

Thomas, M., Gilson, R., Ziulowski, S. and Gibbons, S. (1989) Short term memory demands in processing synthetic speech **Proceedings Human Factors Society 33rd. Annual Meeting** Santa Monica, CA: Human Factors Society 239-241

Thompson, B.H. (1980) Linguistic analysis of natural language with computers **Proceedings 8th. International Conference on Computational Linguistics**

Thorndyke, P.W. (1977) Cognitive structures in comprehension and memory of narrative discourse **Cognitive Psychology 9** 77-110

Tomita, M. (1986) An Efficient Word Lattice Parsing Algoritm for Continuous Speech Recogntion **IEEE Proceedings International Conference ASSP** 1569-1572

Treisman, A.M. (1960) Contextual cues in selective listening **Quarterly Journal of Experimental Psychology 12** 242-248

Treisman, A.M. (1964) Verbal cues, language, and meaning in selective attention **American Journal of Psychology 77** 206-219

Umbers, I.G. (1979) Models of the process operator **International Journal of Man Machine Studies 11** 263-

Underwood, G. (1974) Moray vs. the rest: the effects of extended shadowing practice **Quarterly Journal of Experimental Psychology 26** 368-372

Underwood, M. J. (1977) Machines that understand speech **The Radio and Electronic Engineer 47 (8/9)** 368-376

Underwood, M.J. (1980) What engineers would like to know from psychologists In Ed. J.C. Simon **Spoken Language Generation and Understanding** Dordrecht: Reidel

Usher, D.M. (1986) **A Software Package for the SR128 Speech Recogniser** Bristol: C.E.G.B. Report Number SWR/SSD/0782/N/86

Usher, D.M. (1988) **The Macrospeak Speech Recogniser** Bristol: C.E.G.B. Report Number: OED/STM/88/10081/N

Usher, D.M. and Baber, C. (1989) **Automatic Speech Recognition in the Grid Control Room: Part 2. Development and Assessment of a Telecommand Demonstration** Bristol: National Power Report TD/STM/89/10031/N

Vallar, G. and Baddeley, A.D. (1984) Phonolgical short term store, phonological processing, and sentence comprehension: a neuropsychological case study **Cognitive Neuropsychology 1** 121-141

van Dijk, T.A. (1984) Dialogue and Cognition In Ed. L. Vaina and J. Hintikka **Cognitive Constraints on Communication: Representations and Processes** Dordrecht : Reidel

van Dijk, T.A. and Kintsch, W. (1983) **Strategies in Discourse Comprehension** London: Academic Press

van Cott, H.P. and Kinkade, R.G. (1972) **Human Engineering Guide to Equipment Design** New York: Mc.Graw Hill

van Nes, F.L. (1986) Human factors engineering of interfaces for speech and text in the office **IPO Annual Progress Report 21** 88-94

Vaughan, J. Brookes, G., Chalmers, D. and Walts, M. (1987) Transputer application to speech recognition **Microprocessors and Microsystems 11(7)** 377-382

Vicens, P. (1969) **Aspects of Speech Control by Computer** Unpublished PhD Thesis Stanford University

Vidulich, M.A. (1988) Speech responses and dual task performance: better time sharing or asymmetric transfer? **Human Factors 30(4)** 517-529

Viglione, S. (1986) Speech Recognition: Consumer Products Voice Processing - **The New Revolution Proceedings of the International Conference on Voice Processing** Pinner: Online

Visick, D., Johnson, P. and Long, J. (1984) The use of simple speech recognisers in industrial applications In Ed. B. Shackel **Interact '84** Amsterdam: Elsevier

Walker, R.E., Nicolay, R.C., and Stearns, C.R. (1965) Comparative accuracy of recognising American and International road signs **Journal of Applied Psychology** 49 322-325

Warren, C. and Morton, J. (1982) The effects of priming on picture recognition **British Journal of Psychology** 73 117-13

Waterworth, J.A. (1984) Speech Communication: How to Use it In Ed. A. Monk **Fundamentals of Human Computer interaction** London: Academic Press

Waterworth, J.A. (1987) **Speech and Language Based Interaction with Machines: Towards the Conversational Computer** Chichester: Ellis Horwood

Waterworth, J.A. and Holmes, W.J. (1986) Understanding machine speech **Current Psychological Research and Reviews** 5 228-245

Weiner, N. (1948) **Cybernetics, or Control and Communication in the Animal and the Machine** New York: Wiley

Weizebaum, J. (1966) ELIZA - a computer program for the study of natural language communication between man and machine **Commuincations of the ACM** 9 36-45

Welch, J.R. (1977) **Automatic Data Entry Analysis** Rome, NY: Rome Air Development Centre Report RADC TR - 77 - 306

Wetterlind, P. and Johnston, W.L. (1987) An emergency command recogniser for voiced system control **Proceedings of 24th. Annual Symposium SAFE Association** 181-184

Wever, R. A. (1988) Circadian control of vigilance In Ed. J.P. Leonard **Vigilance: Methods, Models and Regulation** Frankfurt am Main: Peter Lang

Wheddon, C. and Linggard, R. (1990) **Speech and Language Processing** London: Chapman and Hall

Wickens, C.D. (1980) The structure of attentional resources In Ed. R. Nickerson **Attention and Performance VIII** Hillsdale, NJ: Lawrence Erlbaum

Wickens, C.D. (1984) **Engineering Psychology and Human Performance** Columbus, Ohio: Charles E. Merrill

Wickens, C.D., Sandry, D.L. and Vidulich, M. (1983) Compatibility and Resource competition between modalities of input, central processing, and output **Human Factors** 25 227-240

Wickens, C.D.,Vidulich, M. andSandry Garza, D. (1984) Principles of S-C-R compatibility with spatial and verbal task: the role of display control location and voice interactive dispaly control interfacing **Human Factors** 26 533-542

Wickens, C. D. and Weingartner, A. (1985) Process control monitoring: the effect of spatial and verbal ability and concurrent task demand In Ed. R.E. Eberts and C.G. Eberts **Trends in Ergonomics/Human Factors II** Amsterdam: North Holland

Wickens, C.D. and Liu,Y. (1988) Codes and Modalities for multiple resources: a success and some qualifications **Human Factors** 30 599-616

Williams, C.E. and Stevens, K.N. (1972) Emotions and speech: some acoustical correlates **Journal of the Acoustical Society of America** 52 (4) 1238-1250

Williamson, D.T. and Curry, D.G. (1984) Speech Recogniser Performance Evaluation in Simulated Cockpit Noise **Proceedings of SpeechTech'84** 99-102

Wilensky, R. G. (1978) Why John married Mary: understanding stories with involoving recurring goals **Cognitive Science** 2 235-266

Williges, B. H. and Williges, R. C. (1982) Structuring human computer dialog using speech technology In Ed. D.S. Pallett **Proceedings of the Workshop on Standardisation of Speech I/O Technology** Gaitherzburg, MD: National Bureau of Standards

Williges, B.H., Schurick, J.M., Spine, T.M. and Hakkinen, M.T. (1986) Using speech in the human computer interface In Ed. R.W. Ehrich and R.C. Williges **Human Computer Dialogue Design** Amsterdam: Elsevier

Wilpon, J.G. and Roberts, L.A. (1986) The Effects of Instructions and Feedback on Speaker Consistency for ASR **IEE Conerence on Speech I/O: Techniques and Applications** London: IEE Pub. no. 258 242-247

Wilson, J. (1986) Applications of speech recognition in industrial and military environments **Proceedings Voice Conference - the New Revolution** Pinner: Online

Wilson, M.D., Barnard, P.J., Green, T.R. and McLean, A. (1988) Knowledge based task analysis for human computer systems In Ed. G.G. van der Veer et al. **Working with Computers** London: Academic Press

Winograd, T. (1972) **Understanding Natural Language** New York: Academic Press

Winograd, T. and Flores, C.F. (1986) **Understanding Computers and Cognition: a New Foundation for Design** Reading, Mass: Addison Wesley

Wiren, J. and Stubbs, H.L. (1956) Electronic binary selection system for phoneme classification **Journal of the Acoustical Society of America** 28 1082-1091

Witten, I.H. (1982) **Principles of Computer Speech** New York: Academic Press

Wittgenstein, L. (1958) **Philosophical Investigations** London: Basil Blackwell

Woods, D.D. (1984) Some results on operator performance in emergency events In Ed. D. Whitfield **Ergonomic Problems in Process Control** London: Institute of Chemical Engineers

Woods, D.D., O'Brien, J.F., and Hanes, L.F. (1987) Human factors challenges in process control: the case of nuclear power plants In Ed. G. Salvendy **Handbook of Human Factors** Chichester: Wiley

Woods, D.D. (1988) Coping with complexity: the psychology of human behaviour in complex systems In Ed. L.P. Goodstein et al **Tasks, Errors and Mental Models** London: Taylor and Francis

Yannakoudakis, E.J., and Hutton, P.J. (1987) **Speech Synthesis and Recognition Systems** Chichester: Ellis Horwood

Yellowpages (1990) Personal communication regarding telesales staff

Yerkes, R.M. and Dodson, J.D. (1908) The relation of strength of stimulus to rapidity of habit formation **Journal of Comparative and Neurological Psychology 18** 459-482

Young, A.W, McWeeny, K.H., Ellis, A.W., and Hay, D.C. (1986) Naming and categorising faces and written names **Quarterly Journal of Experimental Psychology** 38a 297-318

Zarembo, C.A. (1986) The usefulness of voice recognition on the floor of the New York stock exchange **Proceedings of SpeechTech'86** New York: Media Dimensions 69-72

Zoltan, E., Weeks, G.D., and Ford, W.R. (1982) Natural language communication with computers: a comparison of voice and keyboard inputs In Ed. G. Johannsen and J.E. Rijnsdorp **IFAC Analysis, Design and Evaluation of Man Machine Systems** 255-260

Zoltan Ford, E. (1984) Reducing variability in natural-language interactions with machines **Proceedings 28th Annual Meeting of the Human Factors Society** Santa Monica, CA : Human Factors Society 768-772