An Assessment of the Contribution of Speech Interfaces to Open Access Education

Presented at the CAOAE Workshop,
Napier College, Edinburgh, July 1990

Technical Report No. 114

Jill Hewitt
Christine Cheepen

November 1990

An Assessment of the Contribution of Speech Interfaces to Open Access Education

Jill Hewitt & Christine Cheepen
ISDIP, Hatfield Polytechnic

Abstract

At Hatfield we have had considerable experience in dealing with blind and disabled students, and since 1988 we have been researching into speech interfaces, paying particular attention to their usability and suitability for different applications and different classes of user.

There are currently an increasing number of commercially available products offering speech input for disabled users and speech output for the blind. Although this is an important advance in improving access to education for the disabled, there are still a number of problems as yet unresolved which militate against the full use of such technology by many disabled users. In many cases improvements to the human-machine interface which are designed primarily with the able bodied in mind make the system harder to use for the disabled. Interface design is one of the major problems we are addressing within the Intelligent Speech Driven Interface Project - 'ISDIP'.

The ultimate aim of ISDIP is to develop a voice interface for the full range of computer applications. We currently have two demonstrators - a speech input word processor and a diary management system using speech output. Experiments are being carried out at Hatfield with disabled and able-bodied users to evaluate the effectiveness of the interface to the word processor and the usability of the system. Major results of these evaluations are reported in this paper.

1. Introduction - computing resources for the disabled

Over recent years considerable progress has been made in providing computing resources for the disabled, to facilitate both their education and their ability to be effective in the work place, and the value of computers to the disabled population is now widely accepted. With appropriately designed computing technology the disabled can now undertake college courses alongside the able bodied and work from home (1), and the Chatback project (2) provides facilities for children with

communication disabilities to participate in dialogues with others throughout the world by electronic mail.

2. Interface design - some problems

The simple provision of a computer work station is not, however, always the whole answer to the communication problems experienced by the disabled. Whatever system is provided must be *usable* (3) - i.e. the input and output media must be suitable to cope with the requirements which are determined by the user's particular disability. For certain classes of disability the standard input/output media of keyboard and screen are at best less than perfect and at worst totally unusable. In many of these cases the best (sometimes the only) kind of computer access is speech - the blind or partially sighted requiring speech output, and the manually disabled requiring speech input. For some severely disabled users (e.g. blind and manually disabled, or disabled and bed-ridden) it may be desirable to have an interface offering both input and output by speech.

At Hatfield Polytechnic we have had considerable experience in dealing with blind and disabled students (4). Drawing on that experience, this paper reviews the requirements for speech based interfaces, making particular reference to the ongoing research at Hatfield in the Intelligent Speech Driven Interface Project (ISDIP).

3. Interface design - some solutions

3.1. Speech input

Among the many computing applications for which it is desirable to provide speech input, the most generally useful is the word processor, and several speech input word processors are now commercially available. Most of these work on the principle of recognition of isolated words, and all of them are speaker dependent, and therefore require to be trained by each individual user. This usually involves the user in providing a sample (in most systems several repeat samples) of all the words (items) he/she wants the system to recognise. The system then stores these samples as templates, and matches incoming words when the system is in use against the stored templates. The total number of words can be up to around 30,000, but in order to cut down the system search time when matching templates these must be stored in smaller vocabularies which can be swapped in and out during a usage session.

Different systems offer different sizes of vocabulary, ranging from 50 to 500 words immediately accessible for searching at any one time. Some experimental speech systems, employing phoneme recognition as opposed to word recognition, now have huge vocabularies (the IBM Tangora system claims 20,000 words), but this necessarily creates problems with recognition accuracy (IBM state that the error rate on a 'typical memo' created by Tangora is 6.3 per cent).

While the currently available systems have a high degree of accuracy (most claim 97 per cent or more) and offer a large total vocabulary, they are often designed primarily not to provide speech as an *alternative* to the keyboard, but to provide the facility for speech to be used *with* the keyboard. The kind of interface they typically provide does not take into account the special requirements which arise when the only possible input medium for a user is speech, and consequently their potential for use by the disabled is severely limited.

Within ISDIP we have developed a speech input word processor with an interface which is specifically designed for voice input as an alternative to the keyboard, bearing in mind the fact that for many disabled people the use of the keyboard is either impossible or unreasonably demanding, given the degree of their disability. In order to do this, we identified three basic problem areas which the interface must be adequate to deal with - *feedback*, *error recovery* and *control*.

- Feedback

  This must be automatically provided by the interface in order to cater for the requirements of the user, e.g.

  - what vocabulary is the system currently searching,
  - if the system misrecognises an item what is the user to do next,

  and is also necessary for what can be thought of as the requirements of the system, e.g.

  - is the user speaking at a suitable volume for the system to recognise the input.

- Error recovery

  In any speech driven system, however accurate, there will inevitably

be occasions when the system misrecognises input items. Efficient strategies must be available to deal with such cases, e.g.

- where the match between input and template is imperfect - the system must check the match with the user, necessitating some kind of human/machine dialogue,
- where a word has been badly trained causing repeated misrecognitions - dynamic retraining must be available,
- where the system is totally unable to recognise a word - a 'failsafe' mode must be provided.

- Control

  There are two fundamental requirements in this category:

  - the system must be usable totally hands-free (this is particularly relevant for disabled users),
  - the speech facility must be capable of being switched off so that the user can go into keyboard mode if desired (again, this is necessary for disabled users, who may be interrupted during a work session and wish to hand on the task to another user).

In addition to these basic requirements, to be optimally efficient the interface should make further provisions to cater for speech as the input medium, but these involve some degree of machine 'intelligence', i.e. they require knowledge of the particular application and (ideally) knowledge of the particular user. They fall into two categories - *error recovery* and *prediction*:

- Error recovery involving machine 'intelligence'

  We have identified two problem areas here:

  - the necessity for an undo/redo command if not provided by the underlying application,
  - the necessity to reduce the potential for 'large' errors (where an undo command is inappropriate) by providing a hierarchy of sub-dialogues to check for accuracy at each stage of the task.

- Prediction

  This should operate in parallel with the basic recognition (pattern matching) mechanism. It involves incorporating into the system the

ability to 'guess' what can be expected to come next at any point in the task in terms of:

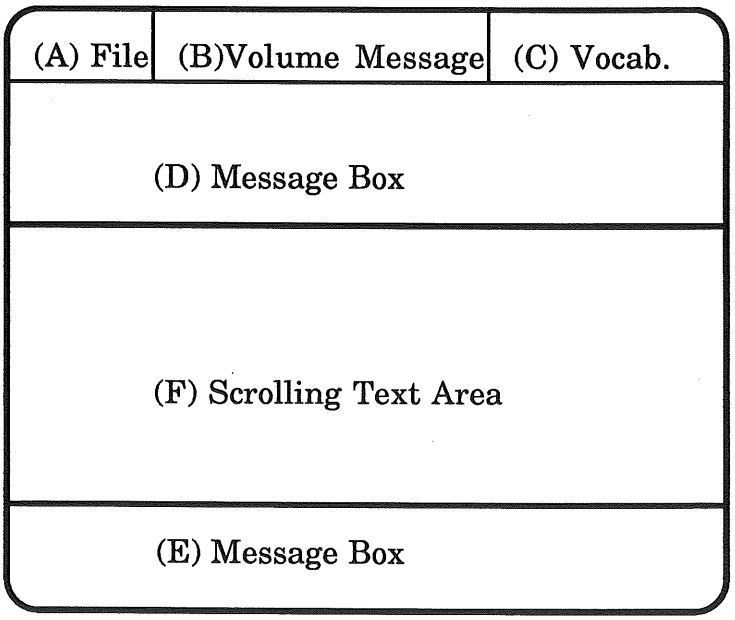
- the syntactic unit currently being constructed
- what the current application will allow
- how the current user normally proceeds with the task in hand.

3.1.1. The ISDIP speech input word processor

The ISDIP word processor is based on the VOTAN single word recognition card, which allows multiple vocabularies of up to 64 words each. Once the word processor has been accessed the system is totally hands free, but provides a facility to switch into and out of keyboard mode. The interface has been designed for speech input, and we have therefore had to address the problems inherent in such a system and to provide for the requirements outlined above. Output is via the screen.

Feedback

As we have said, feedback to the user is essential in any speech system. He/she needs to know a) which vocabulary is currently in use, b) whether the volume of his/her voice is suitable, and c) what to do if the system does not recognise an input item. The ISDIP screen has been designed to provide permanently available display boxes for this information.



The labelled regions are used as follows:

A. The name of the file being edited.
B. Reserved for messages such as "Speak louder please" or "Speak quieter please".
C. The name of the vocabulary currently in use by the recogniser. Where the vocabularies are small this is particularly important, as the user must know when it is necessary to switch vocabularies in order to access a particular item.
D. The main area for error messages.
E. Reserved for sub-dialogues between the user and the interface mangement system and for auxiliary error messages.
F. The scrolling text area.

Pop-up windows appear in the text area for the file menu and the edit menu.

Evaluations

The ISDIP interface as described above is being evaluated by a series of experiments using first time users, both disabled and able bodied (5,6). Volunteers are assessed individually and are not allowed to view sessions preceding their own, so that the system is new to each one. They are given a short demonstration of the training process, followed by a part of a letter being dictated and then saved to disk. The training program is then started for them and they train each word once before embarking on two separate attempts to create a letter, the text of which is provided for them. They have access to a limited number of editing functions and an 'undo' command, and are told to watch the screen for possible error messages while creating the text.

The aim of the experiments is to analyse users' responses to different human-machine dialogue structures and to establish usability guidelines for speech driven dialogue design. Users' responses to various aspects of the interface are observed and analysed; one particular aspect is relevant here, as it most clearly illustrates the importance of feedback, and especially the *kind* of feedback which the user expects when the input medium is speech.

Experiment

The aim is to compare different strategies which can be used when an

input item cannot be perfectly matched with an existing template. These are: a) no error recovery (the system displays the wrong word on the screen) and b) an error message initiated by the machine - that message (which appears in the section of screen marked D on the preceding diagram) simply reads "no match".

In the sample of volunteers who were used for this experiment eight out of nine preferred a message of "no match" to a wrong machine action. This result is particularly interesting as it is unconnected to recognition rate, which in some cases was slightly better for the 'wrong machine action' version (where the number of "undo" commands was higher in order to cope with the wrong actions). Our hypothesis is that the simple error message is preferred because it is consistent with the rules of human-human dialogue, where, if a word is misheard, the listener will give the human equivalent of a "no match" message, such as "pardon" or "can you repeat that please". This gives a valuable indication of the human user's conceptual model of a speech driven interface. Whereas the user may accept 'unnatural', machine-like feedback when using a keyboard, if the input medium is speech, the user will expect feedback which is closer to natural dialogue.

3.2. Speech output

Until recently, systems providing synthetic speech output have been primarily aimed at blind users; however there is an increasing usage of speech output systems at the end of telephone lines - to provide information on directory enquiries for example, or banking or financial services. It is interesting to note that many of these systems use real digitized speech rather than a text-to-speech synthesizer to provide information, even though it is a lot more complicated and memory-consuming. This is because the quality of synthetic speech has been considered too poor to be easily understood at the end of a telephone line. Blind people have however had to put up with systems providing very poor synthetic speech for some time, the assumption being presumably that 'they will get used to it'. In fact there is some evidence for this opinion, in that a survey of blind computer users showed a very positive attitude towards synthetic speech equipment in preference to braille terminals (7), but that is not say that they would not choose good quality speech in preference to poor quality if they were given a choice. It is possible that the strain of understanding poor quality synthetic speech puts an increased cognitive load on the user - more research is needed in this

area.

In addition to the quality of the speech, certain other factors have to be taken into account when designing a speech output interface. The most important are:

- Control.
  The user must be made to feel in control of the interface and be able to repeat phrases, terminate unnecessary dialogues and even get the system to spell words that cannot be understood.

- Feedback.
  The system must provide feedback as to the current mode it is in and feedback on user input. The level of feedback should be adjustable by the user (e.g. should the system echo each character typed, each word or each sentence?).

- User-tunable parameters.
  The acceptability of pitch and speed of the synthetic speech are very subjective and should be adjustable by the user. A common complaint by blind users (7) is that speech synthesizers were not fast enough; most experienced users can understand very fast artificial speech, maybe up to 500 words a minute, although this takes a great deal of practice.

In addition, if a system is to be used by intermittent users it is important to have an acclimatisation phase of relatively unimportant dialogue so the user has time to tune their ear to the particular speech.

3.2.1. The ISDIP Diary Management System

In order to evaluate some of the interface requirements for a synthetic speech output system, we have built a diary management system which takes keyboard input and uses speech output (a later version will use speech input instead of the keyboard). This allows the recording and review of diary entries for an individual or group of individuals. Significant interface features incorporated into this system are:

- user tunable parameters for pitch and speed
- bypass of menu output for experienced users
- easy recall of the last utterance
- generally small menus

- a time-delay prompt which is activated if a response has been made in a specified time

Initial evaluations of this interface with sighted users (with the screen turned off!) and blind users, has proved favourable, but more field studies are necessary to assess its usability and potential. We envisage that a version utilising speech input and output could provide a very useful service at the end of a telephone line for any business person who is out of the office for the day and wants to review their appointments or make new ones.

4. The need for further research

Although there is currently a great deal of laboratory experimentation being carried out on various aspects of speech interfaces there is a great need for more field studies to be carried out on systems in actual use by experienced users. In this way such factors as fatigue and inhibition can be properly assessed; for example, a user may find it relatively easy to create a document using a speech input system in a laboratory, but what would be the effect of using such a system in an office or school environment, and how quickly would the user's voice tire? The disruption caused by a blind user's speech output system can be minimised by providing them with an earphone, but it would be interesting to note to what extent this cuts them off from the 'normal' dialogues with colleagues.

For the specific area of open access education, we would like to see a controlled field study where selected users were analysed over a period of time using particular types of equipment. This would allow us to predict much more accurately the requirements for speech based interfaces. This proposal obviously also extends to other specialist types of equipment and would help considerably in the future design of usable systems for the disabled. We would also recommend that wherever possible disabled researchers are employed to work in this area, as they are much better placed to understand the particular requirements of disabled users.

5. Conclusion

This paper has reviewed some of the requirements for human-computer interfaces utilising speech input and/or output, and has reviewed some of

the progress made by ISDIP in designing usable speech-based interfaces. Speech technology is now at a stage where it can make a real contribution to interfaces for disabled users, and it is to be hoped that many more potential students will be be able to take advantage of a speech-based computer interface to aid them in their studies.

Whereas the severely disabled may of necessity have to work from home, there is a real possibility that speech interfaces will allow more disabled students to study alongside their able-bodied colleagues. Much more funding is needed to cover the still relatively high cost of a speech based system, as is a centralised policy that provides such systems to disabled students as a right rather than as a 'favour'. There is also still a need for educating many educational establishments which tend to view any specialist equipment as a hindrance rather than as an exciting asset.

References:

(1) Teleworking Applications and Potential (TeAPoT). A Feasibility Study of Home-Based and Centre-Based Telework for People with Physical Disabilities. Work Research Centre and National Rehabilitation Board, Dublin 1989.

(2) Chatback Project, Tom Holloway, IBM Support Centre, Warwick.

(3) Hewitt, J.A. & S. Furner, 1988, Text Processing by Speech: Dialogue Design and Usability Issues in the Provision of a System for Disabled Users, in People & Computers IV, Jones, D.M. & R. Winder (eds.)

(4) Hewitt, J.A., 1988, Audio Technology in IT Education for the Blind, presented at Workshop on New Audio Technology for the Blind, Copthorne, West Sussex, October.

(5) Zajicek, M., 1990, Evaluation of a Speech Driven Interface, Proc. IEEE UK IT 1990.

(6) Zajicek, M. & J.A. Hewitt, 1990, An Investigation into the use of Error Recovery Dialogues in a User Interface Management System for Speech Recognition, Interact '90.

(7) Flynn L.J., 1990, Evaluation of Human-Computer Interaction For the Visually Impaired. BSc Project Hatfield Polytechnic.