Proceedings of the Institute of Acoustics

THE COMPUTER AIDED LEARNING ENHANCEMENT PROJECT (CALE)

Christine Cheepen (1) & James Monaghan (2)

(1) and (2)University of Hertfordshire, Speech & Language Technology (*SLANT*), Hatfield, Herts, England, AL10 9AB

1. INTRODUCTION

The Computer Aided Learning Enhancement project (CALE) is being carried out by the University of Hertfordshire multi-disciplinary Speech & Language Technology Team (SLANT), which involves researchers from the disciplines of Computer Science, Linguistics, Education, Business and Engineering. The aim of CALE is to develop a totally hands-free computer workstation for disabled school children, which will allow them to access the full range of standard computer applications by speech input. The work is being partly supported by a substantial financial donation from Save & Prosper Educational Trust.

Our immediate goal is to provide a fully operational system for the children at Lonsdale School for the Disabled, Stevenage, Herts., but in the long term we intend the system to be generally useful in any school where there are disabled pupils who find keyboard input difficult or impossible. For this reason, one of our highest priorities is to ensure that the overall cost of the system is kept as low as possible, so that it will not overstrain a limited school budget, and we are using standard, typical school hardware, so that no extra cost is imposed by having to purchase very expensive machines. For the Lonsdale development, we are using a Nimbus 386 platform, as there is already a machine identical to this in the school. We are currently carrying out a survey of schools in the area (Hertfordshire, Bedfordshire, Buckinghamshire, parts of North London), in order to get firm information as to which machines are in most common use.

2. PROJECT DESIGN

2.1 Target population.

Lonsdale School currently has around 75 pupils, their ages ranging from 3 to 18, so it is a complete school environment, from intake at nursery school level up to 6th form. All the children have some form of physical disability, and their problems include blindness, deafness, various kinds of paralysis, cerebral palsy and muscular dystrophy. In addition to this, all the children have some degree of learning difficulty, very pronounced in some cases, less noticeable in others.

2.2 Basic approach

Our experience in previous projects has shown that, when designing systems for end users, it is essential to involve a sample of those end users early on in the design process, and to regularly consult them while development is in progress. This is vital if the final system is to be really usable. In the case of disabled end users it is even more important to keep them involved throughout the design and development phases, as the perception of a person actually using a system over a lengthy period to carry out his work can be very different from

the perception of the designer who may only test out the development in short sessions, to ensure that the implementations are functioning properly, and it is extremely difficult (if not impossible) for an able bodied designer to put himself in the position of a disabled end user, who cannot - in the case of a speech input system - get out of difficulties by using other media, such as the keyboard.

We decided, then, to design the CALE project on an incremental prototyping basis, so that as each phase of the design and development is completed it is passed to the end users at Lonsdale school, who then test it out over a period of weeks, without supervision by the design team, and report back their evaluations for inclusion in the next version of the prototype. This approach is particularly appropriate in a project such as CALE, which is a two year development, ending in April 1994. The problem with user reactions to a system is that very often an inexperienced user will have very different perceptions from an experienced user. What is, for the novice, a helpful (maybe an essential) aspect of the human-computer interface, can be irritating for the experienced user, and can slow down his interaction with the system, perhaps by offering him too much help which he no longer needs. The continuing feedback from the Lonsdale end users will provide us with information across this spectrum, and this will be a central consideration in the developing design, as our intention is to produce a final system which is sufficiently adaptable and flexible to be maximally usable by both the first time user and the expert.

2.3 Scope of the project

With a large range of disabilities such as exist at Lonsdale it is not practical to try to produce one system which will solve all problems, and we have therefore restricted ourselves to dealing with a subset of the pupils - those who have difficulty in using the keyboard, but are able to see well enough to read the screen. Some of this group (approximately 15 children) have some hand movement, and are able to use the keyboard to some extent or to use a switch system. This is not, however, really suitable for them, as the effort which is required to operate this kind of input medium is considerable, and they tire very quickly. An additional problem is that for those with muscular dystrophy their ability to use their hands diminishes with time as the disease progresses. In order to cater for these children, and also those who are totally unable to use their hands, we decided on a speech input system, which would allow them to access, by voice input, all the standard educational software packages which they need for their school work. Clearly the CALE system, which allows speech input but outputs text to the screen is not suitable for blind users, but there is no reason why this cannot, at a later date, be combined with a text output system (perhaps a simple screen reader) in order to accommodate the needs of blind or partially sighted users.

Our ultimate goal is to provide a system which will allow the users complete independence from able bodied help, and to achieve this, the CALE team members whose expertise lies in mechanical, electrical and electronic engineering are, in parallel with the work on the speech input core of the system, developing a wheelchair operated switch, to allow the children to turn the machine on and off. This will ensure that they have full independence, and are not reliant on staff or helpers when using the system. The wheelchair operated switch will have substantial advantages over existing switches for the disabled, which tend to be based on suck/blow devices and the like, and are very overtly *specially for the disabled.* Our system will provide a more 'normal' way of operating, which will have psychological benefits for the

users over and above the basic functionality.

3. THE CALE SYSTEM

3.1 Fundamental design and development decisions

In order to allow full voice access to a wide range of software applications it is necessary to have a large, flexible vocabulary, particularly for coping with tasks such as word processing - a pressing requirement at Lonsdale. This meant that, if we were to take full advantage of the currently available speech recognition technology, we were constrained to produce a speaker dependent system (workable speaker independent systems being far too restricted in vocabulary size). As an important consideration was cost, so that future systems would be affordable within a school budget, it was also necessary to ensure that the core recognition system was as inexpensive as possible, while still providing a vocabulary large enough to cope with potential user requirements across a range of standard computer packages.

With these constraints in mind we investigated the commercially available systems, and also considered our own, in-house prototypes which we developed during the Intelligent Speech Driven Interface Project (ISDIP) - a speech driven word processor and a generalised speech interface for use with simple commercial pc packages.

In terms of simplicity and ease of use, which is very important for a novice user, our in-house word processor has great advantages, as it was designed specifically for use by voice alone, and has very clear, simple to operate strategies for dealing with problems caused by misrecognitions. It does, however, have problems which arise from the constraints on the underlying speech recognition card - a Votan vpc200. This imposes sectioning on the total vocabulary, so that the recogniser can only address one section at a time. On the old card the sections were up to 64 words each, and on the latest version they are 128 words each. The user has responsibility for switching between vocabularies to find the required item, and this becomes a problem when the total vocabulary is very large, as the user must remember the names of the different vocabularies and how they are structured - i.e. which items are stored where. In a large vocabulary task such as word processing, the cognitive load which this imposes is unacceptably high. On the other hand, the robustness and reliable recognition offered by the Votan card is an advantage, particularly when dealing with users who may become frustrated if their work is constantly being interrupted by imperfect recognition of their input.

Our in-house generalised speech interface - also based on the Votan card - although suitable for speech input to very simple commercial packages, such as games, databases, or spreadsheets, where the required vocabulary is very restricted, is not sufficiently robust or flexible to be used across the board with the full range of computer applications. Again, however, it has an advantage over some commercial systems in that it is very simple to use.

We decided, on balance, that for the pupils who are capable of using school software packages to their full capacity, our in-house developments were unsuitable, as they did not provide sufficient flexibility or functionality to cope with the users' needs. We therefore decided to base our CALE system on the IBM VoiceType, which is a generalised speech

interface, offering full functionality across the range of commercially available pc software packages which are normally accessed via the keyboard. It allows a total vocabulary of 7000 words - 5000 (including command words and a large dictionary) provided with the system, and a further 2000 which can be chosen by the user, and altered as desired. VoiceType is an extremely powerful system, which is, inevitably, rather complex to operate, particularly for a naive user. Bearing in mind that our immediate target population at Lonsdale are all children with learning difficulties, we were aware at the outset that our choice of VoiceType would inevitably involve us in a considerable amount of customisation, in order to ensure that the final system would appear sufficiently simple to be easily operable by the children.

In the case of the children with very severe learning difficulties, we realised that the full functionality offered by VoiceType would not, in practice, be required, as the children would be unable to exploit the full potential of the educational software with which it can interface. For those children we are providing our in-house word processor, with a very limited vocabulary, so that they will be enabled to produce very simple texts using voice as their input medium. This compromise solution, of installing two different systems at Lonsdale will have advantages in terms of fundamental research into speech interfaces, as both systems will be accessible by the more mentally able children. Their evaluations of the two systems in parallel will allow us to draw conclusions about interface design, and may provide valuable insights into the often problematical area of the trade-off between simplicity and functionality.

3.2 VoiceType and the CALE system - advantages
Using VoiceType as a core recognition system has four major advantages for the final CALE system - large total vocabulary, flexibility, 'concealed training', and affordability.

3.2.1 Large total vocabulary. Unlike earlier commercial products, VoiceType provides a large vocabulary (5000 VoiceType words, including commands and a large dictionary, plus 2000 user-chosen words which can be altered at will), but does not require the vocabulary to be sectioned. This is an important point for the user, as it removes the need to remember where particular items are stored. Much of this enhanced usability is due to the fact that, unlike the Votan card (and many other products), which works by simple template matching of the input signal with a previously trained sound pattern for the stored item, it uses Hidden Markov Modelling to build models of speech segments, takes the incoming signal as a collection of these models, and translates it into the nearest orthographic form offered by the on-line dictionary (which includes the user chosen words). From the user's point of view, this simplifies the business of recognition greatly, as he simply says a word and the system produces an output item for it, without the user being aware of how this item is stored.

3.2.2 Flexibility. This is an important consideration in a system designed for school use. While word processing is the most pressing need for the children at Lonsdale, the range of school work they can attempt unaided is greatly increased if they can access a system which will also allow them to use databases, spreadsheets and whatever other packages are educationally useful, in addition to quite demanding games such as chess.

3.2.3 'Concealed training'. With a very large vocabulary system such as VoiceType, this is an essential. Unlike the Votan card, which requires the user to train each vocabulary item only once (in itself a very time consuming task when the vocabulary is large), VoiceType requires three instances of the *basic* vocabulary in order to build the voice models. The basic

vocabulary - i.e. most of the important command words - is 200 words, and this training process obviously takes some considerable time. However, it is the only *overt* training the user is required to do. All other vocabulary items, whether provided by VoiceType or the user, are trained *as they are used.* On the first occasion a user inputs an item of his own he is required to say it, then spell it (using the pilot's alphabet). At this point it is automatically added to the dictionary, together with the model created when he said the item. As he uses the item more often, the voice model is updated (as with all VoiceType words and commands), so that each time the system is used functions to update the training on every item input by the user. The user is not, therefore, really aware of having to 'train the system' - his perception is, rather, that the system is 'getting to know him', which adds to user satisfaction when interacting with the system.

3.2.4. Affordability. At present, the VoiceType system costs around £3000, and comes complete with software, headset and microphone. This is not a trivial sum, but, in comparison with other systems, is 'affordable'. It seems likely that as the system becomes more widely used (it has only recently been introduced into the UK), the price should fall, so as to bring it more directly within the scope of the standard school budget.

3.3 VoiceType problems and some CALE solutions
Along with its many advantages, VoiceType has its problems. These fall into three areas - limitations, complexity, and feedback.

3.3.1 Limitations. One of the limitations of VoiceType is the restriction of the vocabulary to 7000 words in total. As we have already explained, this means 5000 words provided by the system at the outset, plus another 2000 which can be chosen by the user. If the user puts in more than 2000 then the least recently used item will be deleted from the dictionary. Although it is possible to envisage a scenario in which this restriction could be a problem for the user, in the context of the CALE project we do not anticipate any difficulties. Firstly, our judgement is that a vocabulary of 7000 words will be ample for our target population (in fact it would be ample for most populations in most domains). Secondly, whereas some of the end users may find that they eventually do lose a previously input item, due to adding more than 2000 words, this will not be an item which they require very often, as VoiceType always deletes the item which has been used least recently. If the user's vocabulary requirements change and the word becomes useful again, then he can input it as a new item, and VoiceType will update the usage log, so that it will become a recently used item, which will not therefore be deleted to make room for further additions.

Another limitation is that VoiceType is an isolated word recognition system. For users who are unfamiliar with speech recognition devices, this often creates problems at the outset, as their expectation is that if the machine will recognise their voice they can therefore chat to it as they would to a human being, and they expect to be able to talk at a normal conversational speed. Although there are some systems at present available which claim to recognise continuous speech, the CALE design and development team have yet to see one which lives up to those claims. The closest we have seen simply recognises phrases, such as "thank you very much" or "four door Ford Sierra". This is NOT continuous speech recognition in the true sense, in that such systems cannot cope with connections which have not been explicity trained (such as the audible "r" in "here it is", which is not, in most British accents, audible in

the isolated word "here". The phrases which are recognised by such systems are simply equivalent to long words, and are recognised in isolation from one another. The limitation of isolated word recognition is, then, a fact of life which must be lived with. In practice, where the end users are not copy typing, but composing directly on to the machine, this limitation is not really problematical, as response time for the input item is very short (approx. 1/2 second), and the user is inevitably taking some time to mentally process the input before he utters it. When the CALE project is near completion, we plan to do some time tests, comparing user performance while composing text across a range of input media, including keyboard, speech system and other aids for the disabled, such as mouth sticks and head wands. Differences between overall performance times will be reported in later academic publications.

3.3.2 Complexity. This is a problem which does have an impact on the CALE project, because the flexibility of VoiceType inevitably means that there are a large number of commands which must be used to carry out all the possible actions. Learning these commands takes a considerable amount of time, although the requirement to train the basic vocabulary three times for each item is a help here, as reading every item in the list three times helps to familiarise the new user with what is available.

Our target population at Lonsdale is such that a large part of our effort in the CALE project is devoted to customising the VoiceType core by simplifying it, and concealing much of its power, so that it will be usable by young users who have learning difficulties. We are doing this at present by setting up a series of macros, so that access to different packages in different files or directories will be by simple voice commands which conceal from the user the various steps which are taken to get from one place to another, in and out of packages etc. In addition to this we are also setting up initial vocabularies specially for the Lonsdale children, as this will reduce the potential for error which could arise if they do this completely unaided - bad spellers may input items with incorrect keystrokes. We are also considering adding a spell checker into the system, which would prevent (or correct) problems of the kind in the future. A report on this, including any difficulties encountered will be presented in a subsequent publication. This process of customisation will continue throughout the project, fed continually by the evaluations of the end users as they test each version of the prototype.

3.3.3 Feedback. This is a very serious problem, and one which the CALE design and development team encountered early on in the first weeks of the project. VoiceType is a speaker dependent system, and where a system is set up for more than one user, the voice files must be kept separate, so that one user's voice does not corrupt the files of another. In our first experiments with VoiceType we quickly discovered that this is only too easy to do, as VoiceType allows the user to switch between sets of voice files (so that one user can take over from another without rebooting the system), but it provides no permanent visual feedback to remind the user *who the system thinks is doing the talking*.

As the CALE system is being designed to be used by several children on the same machine, this is a problem we are concentrating on. At present our plan is to alter the prompt to indicate the current user, as the corruption of a user's voice files can be very serious indeed, possibly involving the user in completely starting afresh, with initial training, and then building up his own personal vocabulary from scratch.

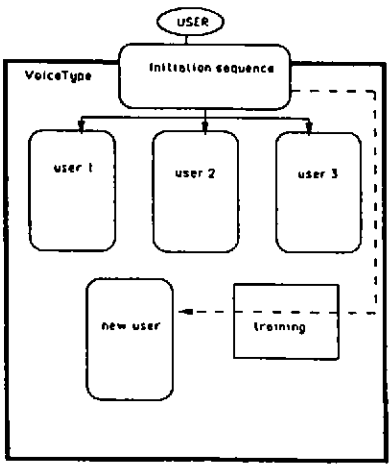
4. SPEAKER INDEPENDENCE IN THE CALE SYSTEM

Although the speaker dependent nature of the VoiceType system is perfectly suitable for use by our target population in a general sense, because our end users are disabled, and cannot use the keyboard, it will be most helpful for them (and essential if they are to be independent of able bodied help) if there is an opening, initiation sequence offered by CALE which is speaker *in*dependent. In this way, they will be able to wheel up to the system, turning on the machine by the wheelchair operated switch as they arrive, and get through to their own voice and application files purely by voice, without needing the help of a keyboard operator. For this, the initiation sequence must recognise whichever user is talking, and one of our major concerns is to develop this initiation sequence.

The vital point here is to keep the vocabulary of the initiation sequence very small, and to ensure that each item in it is easily distinguishable. In this way the system will not misrecognise an item simply because one user's voice is different from another's. This can only be successfully accomplished by setting up an appropriate dialogue for the user and the system, which is initiated by the system, and in which the user is instructed to use only certain responses. Our intention is to design part of it as a menu, so that the user simply has to say a digit in order to choose an option.

Work on this dialogue has already begun, but it will require considerable further effort and extensive testing by the target population before the best implementation can be identified. Our plan is to structure the final system as in fig. 1 below.

Fig. 1



As the diagram indicates, if a user has already acquired voice files, and is therefore a 'known user', he will, via the initiation sequence, be automatically transferred into his own area, with his own files operational, ready for work. If he is a new user, he will, again via the initiation sequence, be automatically taken through the initial training procedure, during which new voice files will be set up for him. When he next accesses the system, he will be a known user and on identifying himself will be transferred directly to his own voice files.

5. BIBLIOGRAPHY

Cheepen, C., 1990, "The pragmatics of friendliness and user-friendliness", **International Pragmatics Conference**, Barcelona, July

Cheepen, C. & J.Monaghan, 1992 "The Computer Aided Learning Enhancement Project", **The International Clinical Phonetics and Linguistics Association Second Symposium**, London, July

Hewitt, J. & C.Cheepen, 1990, "An assessment of the contribution of speech interfaces to open access education", Workshop on Open Access Education, Napier Polytechnic, July

Monaghan, J., 1990, "The pragmatics of human-computer interaction: an overview", **International Pragmatics Conference**, Barcelona, July

Monaghan, J., C.Cheepen, J.Hewitt & J.Hobson, 1990, "Simulation in the evolution of an intelligent speech interface", in Life M A, C Narborough-Hall & W Hamilton (eds), **Simulation and the user interface**, Taylor & Francis

Nusbaum, H.C., 1986, "Human factors considerations in the design of large vocabulary speech recognition devices", **Speech Technology '86**