Introducing Voice Control - Widening the Perspective

Ian French, Philip Halford*, Jill Hewitt, John Sapsford-Francis

University of Hertfordshire, Hatfield, Herts, England, AL10 9AB
Telephone: 44 707 284766 Fax: 44 707 284303 email: comrirf@herts.ac.uk

*Compris Consulting Ltd, Studio II, 2 Beverley Gardens, London, HA7 2AB.
Telephone: + 44 (0) 81 424 2477 Fax: + 44 (0) 81 424 2479

Abstract. This paper describes the development of a multimedia tutorial system which is designed to encourage more people to use voice controlled systems. A three part tutorial takes a user through a first introduction at an exhibition or assessment centre, an intermediate learning level designed to improve performance in using speech controlled systems and finally an application oriented tutorial designed to accelerate learning to 'expert user' status. The system is designed to be used in hands-free mode right from the first access by a new user, thus giving the disabled user more independence of use throughout their training, hence minimising the need for third party assistance.

Keywords. Speech Controlled Systems, Hands-Free Operation, Tutorial System, Multimedia Tutorial.

1 Introduction

The SHELVS project (Self Help Learning for Voice Controlled Systems) is jointly funded by ESRC and the DTI. This is a collaborative project with partners from academia and industry which aims to fulfil the training needs of many people with disabilities who could benefit from the use of voice controlled systems.

There are currently a wide range of commercial speech recognition systems which provide an alternative to keyboard input, but usability problems with these systems mean that they do not attract as many users as they could and that the users often do not exploit the full potential of the systems. Most systems require some keyboard input during the tutorial and training stages, and often have error recovery routines, which, although very powerful, are daunting to a new user. In addition they do not address the very real difficulty that many users have in adapting to the idea of talking to a machine, compounded with the extra burden of learning to use a computer and its applications. The SHELVS system addresses these problems by providing a tutorial management system which allows the user to progress from new user to expert status at their own pace and under their own voice control, thus doing away with the need for expensive consultancy and able-bodied assistance.

In the following sections, we discuss the rationale behind the system, including the results of a questionnaire of existing users, the overall system design and the evaluation programme.

2 Rationale

Commercial speech recognition systems such as the DragonDictate and its derivative the IBM Voicetype record impressive data entry rates of up to 60 words per minute and recognition rates of up to 97% correct (Baker, 1989), and can be operated by experts without recourse to the keyboard. However it is apparent that, particularly for new users, there are still significant usability problems. New users have to undergo a fairly lengthy training period, and typically require the services of a consultant for at least a half to one day before they can begin to use the systems on their own. They need to learn the International Communications Alphabet for when they have to spell words, they need to know how to recover from errors and they need to understand how to build and use macros (single word commands that instigate a series of actions or keystrokes). Typically before they can begin to use a system such as DragonDictate they will have to train and remember 204 command words. All of these factors can be seen as a considerable disincentive to a disabled user who is assessing the relative merits of a voice controlled system against the more traditional switch input device, even though the eventual performance rate of the speech system will outstrip other input modalities.

We set up a simple trial in which an expert, a novice and a new user of IBM Voicetype all tried to input the same text. The expert was a system trainer, the novice had undergone the initial training and had learned the International Communications Alphabet and trained all the command words, and the new user had trained some command words and had some knowledge of error recovery procedures gained through watching other users. The results are given below:

> **Expert:** 59 utterances created 120 words of text (using macros). 4 errors, no use of keyboard, 38 words per minute, recognition rate of 91%
> **Novice:** 42 utterances created 28 words of text, 14 errors, two uses of keyboard, 8.5 words per minute, recognition rate of 67%
> **New User:** 44 utterances, created 15 words of text, 29 errors, three uses of keyboard, one word left wrong, 3 words per minute, recognition rate of 34%

In our opinion, the expert was inputting text at the maximum rate possible for the system (operating on a standard 486 compatible PC). She could only have achieved a better data entry rate by the extensive use of macros. The progress of the novice and the new user was slow enough for us to abandon the trial before all the text had been input. In both their cases, it was essential for the keyboard to be used by the experimenter to correct errors caused by the system consistently misrecognising a command word. If they had been on their own and unable to use the keyboard they would have been unable to progress.

This trial reinforced the opinions of the expert regarding the difficulties faced by new users, and serves to show that there is a need for a user centred approach in designing tutorials and applications for voice controlled systems.

3 System Overview

The SHELVS system is being developed through an incremental prototyping approach (Vonk, 1990) with an emphasis on user involvement throughout the design and implementation. Throughout the project we have been building on experience gained in earlier developments of speech based systems. A main focus of the ISDIP project at Hatfield (Tough, 1990) was to provide a word processor that allowed hands-free operation even in compounded error situations. Results from this research point to the need for a good error recovery dialogue commensurate with the user's expectations (Cheepen, 1990) and more closely related to a human to human conversation than a conventional keyboard operation, even when a restricted vocabulary is used (Zajicek & Hewitt, 1990).

The system being built will provide tutorials at three levels:

- Sampler - to introduce and demonstrate voice controlled systems
- Primer - to allow controlled learning of important aspects of voice controlled systems
- Practitioner - to teach effective use of a voice controlled system for a variety of applications

These are described in detail in sections 3.2 to 3.4

3.1 User Requirements

In order to establish the user requirements for the tutorial system we sent out a questionnaire to recent purchasers of DragonDictate and IBM Voicetype systems. We also elicited the opinion of a voice system consultant. Further requirements were identified as a result of the first prototype evaluation and we expect to further refine the requirements specification as a result of future prototype evaluations.

3.1.1 User Questionnaire

A questionnaire was sent out to 48 recent purchasers and 20 responses were received. The results indicated a wide range of usage patterns and ability levels and showed some degree of dissatisfaction with the systems, they are recorded in some detail in (French et al, 1994). Thirteen people reported problems with the system, the most significant being poor word recognition and incompatibility with Windows (all users had DOS versions of the systems). Twelve people suggested improvements to the system, with the most significant being the need for compatibility with Windows, an improvement in speed and a strong preference for English as opposed to American spelling.

3.1.2 Usability Criteria

In keeping with our usual practice in following a user centred design approach, we drew up usability criteria for the three main parts of the system. These took into account the analysis of existing users and the perceived requirements of new users. They provided a framework for the design of the system and its subsequent evaluations. An overview of these criteria is given in the table in Figure 2, they should be considered in conjunction with the system description in the next section, since different criteria are used for different parts of the system.

Criteria	*Sampler*	*Primer*	*Practitioner*
User attitude	*Priority 1*	*Priority 1*	*Priority 2*
Learning Time	*Priority 1*	*Priority 2*	*Priority 2*
Performance	*Priority 1*	*Priority 1*	*Priority 1*
Error Rate	*Priority 1*	*Priority 2*	*Priority 3*
Retention	*Priority 3*	*Priority 2*	*Priority 1*
Flexibility	*Priority 5*	*Priority 3*	*Priority 2*

Figure 2 Usability Criteria and their priorities for the three parts of the system (Priority 1 = high and 5 = low)

3.2 The Sampler System

This has been designed for an interaction of 10-15 minutes duration and is intended to be used in exhibitions and show-rooms. Its purpose is to allow users to explore the potential of voice controlled systems, to allow users to evaluate informally the suitability, for them, of voice controlled systems and to introduce users to the whole tutorial system.

The system starts with an introductory video loop showing how to engage with the system (...put on the headset microphone, with assistance if necessary) and from this the user is able to select speech or keyboard control. If speech is chosen, a selection of six words (<UP>, <DOWN>, <LEFT>, <RIGHT>, <SELECT> and <QUIT>) are trained which is sufficient to take the user through the rest of this level.

Navigation around the sampler is achieved using direct spatial mapping, this is shown in figure 3. When the user utters one of the control words (<UP>, <DOWN>, <LEFT> or <RIGHT>) the input focus is moved to the relevant button and further information about the option they have selected is given in the space adjacent to that button. The button is activated when the user says <SELECT>. Saying <QUIT> at any point will take the user to the top level of the system.

The user is able to choose between two or three introductory videos detailing the functionality of the application, how speech recognition works and examples of (good and bad practice in) speech recognition. These can be controlled by voice.

This level also incorporates different types of games that the user can play under speech control. One of these games is shown in figure 4. This is a simple maze game in which the user moves a ball (top left) through the maze to a man (bottom right) at the end. The control words to move the ball are the same as used in the navigation in the system, but the user is not required to say <SELECT> after each movement. During the game, the user can say <QUIT> to exit from the game. The purpose of this and other games is to give the user a chance to practice the vocabulary they have learned so far, and to establish if speech is a viable medium for them to use (if for example they have a disability that effects the pronunciation or consistency of their speech).

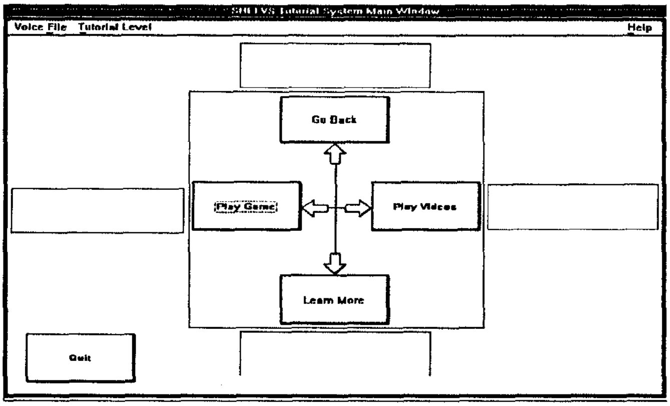


Figure 3 Sampler Screen showing a Directional Spatial Metaphor for selecting system functions

3.3 The Primer System

The Primer system is intended to be used in assessment centres, rehabilitation centres, homes, exhibitions and show-rooms. Its purpose is to allow users to gain the necessary familiarity with a number of speech recognition essentials that apply across the majority of voice controlled systems. These essentials include use of the international communications alphabet to spell words, use of numbers and specific commands. The Primer system also introduces users to environmental control using voice controlled systems.

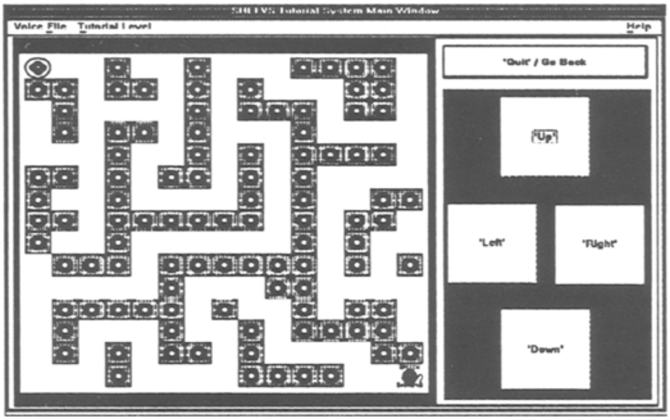


Figure 4 The Maze Game

This part of the system has been designed for a session of about 30 minutes duration, but it is envisaged that users will return for several sessions to reinforce their learning and to cover different aspects of the tutorial. It follows that any words trained by the user during this session will be saved for subsequent use both in this and the Practitioner tutorials. Navigation around this part of the system will be performed in the same way as in the Sampler level. Selection of one of the options at this level will instigate training of the words required.

To allow the user to practice the international communications alphabet a hangman game has been implemented, this is shown in figure 5. The utterence of one of the letters of the alphabet causes the button containing that letter to be disabled. After a couple of tries at guessing words by spelling, the button array shown at the bottom of figure can be made invisible, thus helping the user learn the alphabet.

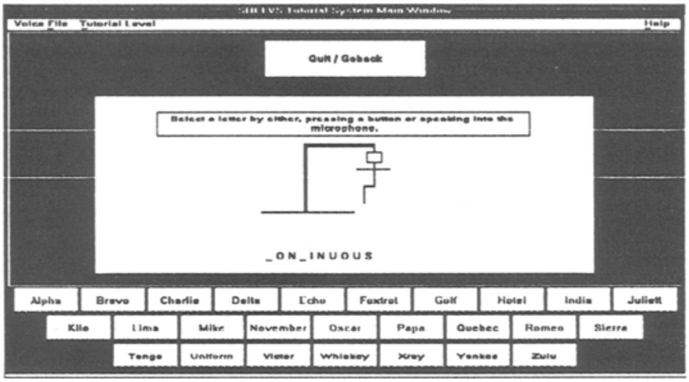
3.4 The Practitioner System

The Practitioner system is intended primarily to be used in the home, but it may also be used in rehabilitation centres and assessment centres. The purpose of this system is to provide users with an orientation to the DragonDictate system and help users to learn the commands and interaction sequences that will allow them to use the system efficiently.

The functionality of the Practitioner will depend to a certain extent on what is provided by DragonDictate with their Windows product (due for release in Spring 1994). The complement of DragonDictate with our own tutorials will provide a complete and usable set of tutorials.

3.5 Dialogue Design

An important principle in the design of the system was that it should be operable by speech even before a user had trained any words. To achieve this end, a "bootstrapping" dialogue has been designed. In addition, the system must cater for existing users who have already trained the initial vocabulary, by loading the appropriate voice files. The Dialogue given in Figure 6 shows how both these requirements have been catered for. The diagram is in USE format (Wasserman et al.)



Select a letter by either, pressing a button or speaking into the microphone.

Figure 5 The Hangman Game

From the start state there is a timeout which passes to a video loop showing how to engage with the system. A message on this screen will tell the user to do anything (keypress, mouseclick or utterance) to engage with the system. The new user then has the option of selecting the Sampler by pressing on a button (taking them to the Message 1 Sampler screen), or by saying "Select Sampler". As they have not yet trained any words, the system will only be able to look for an approximate match for the utterance "Select Sampler" - the degrees of freedom for the recognition will be set very high. Once presented with the Sampler Screen (message2 Sampler in the diagram), the user will be asked to say "Yes" to continue with speech input or "Exit System" to go back to the start. These two words are of sufficiently different length and profile that the system should be able to distinguish between them for most users. A user who has already enrolled with the system will need to establish their identity so that it can load their voice files. They will say a (trained) codename, which, if recognised by the system, will be followed by a request for them to input their password, by spelling one letter at a time. The degrees of freedom on recognition of the codename will be very small to minimise the risk of the system mistaking a new user for an existing one. A further security check is given by the need to input a password, since it would be problematic if a user was allowed access to someone else's voice files.

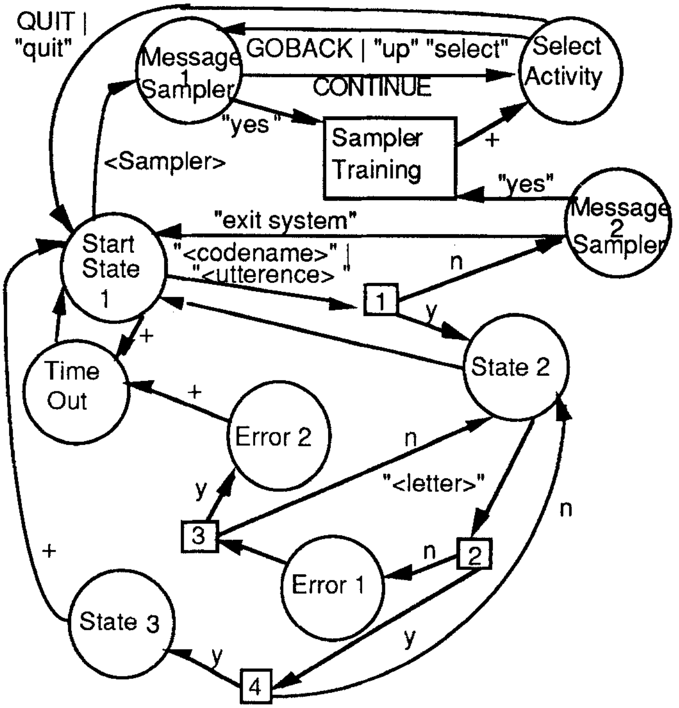
4 Evaluation

In the early stages of the project the development team established links with related product vendors and were able to accomplish two goals to assist the team in the evaluation:

i) to verify the current use profile of the prospective market. i.e. what existing users were using their system for and for how long etc...

ii) to establish a speech-control-fluent user group who were willing to evaluate our products for us.

Prior to field trials of the system, a Hallway and Storefront evaluation was undertaken. This was primarily needed to establish whether the first draft of the system was apt for use by people without the use of keyboard. The team were also keen to see whether the users could use the metaphors adopted in the system to control the application i.e. the sequence of training the utterances and using the games in the first (Sampler) level of the system rather than getting in a total



1 = find match for codename
2 = find match for password
3 = increment failure, count test for > limit
4 = Password complete?
State 1 = "<name>" enter password or stay silent to exit"
State 2 = "Voice files loaded for <name>"
State 3 = "Password Accepted"
Error 1 = "Error, please repeat"
Error 2 = "The system isn't recognising your voice,
please re-enter as a new user or seek assistance"
Message 1 = "Say YES to select speech"
Message 2 = "Say YES to select speech or EXIT SYSTEM to return"

Figure 6 A Bootstrapping Dialogue

muddle. The chosen scenario was a busy canteen area at the University to see how the system would cope with a high level of background noise to simulate an exhibition environment. The team designed a path for the user to take that led the user into the system gently and used a variety of interaction objects supported by MS Windows. This introduced them to a well balanced subset of the system elements. When the users had completed their session with the system, they were de-briefed immediately using a set of screen images that were seen in the interaction, and were asked for their opinion on usability and clarity and for comments on the screen and system layout. In addition to this, the team used a

Mediator system to capture the interaction on video tape discretely rather than obtrusively using an external camera unit. The initial findings from this primary evaluation session were as follows:

- all of the users found the experience using speech positive and were encouraged rather than discouraged to try to use other speech systems
- some of the users found the training of the words confusing
- the initial vocabulary which included <right>,<left> and <quit> was found to be too tight phonetically and led to misrecognitions
- one of the games that was used in the system was considered too difficult to play in the short timescale allocated
- some of the users wanted more on-screen instruction to assist them with the decisions on the screen.

At the moment the team are digesting the corpus of information amassed from this evaluation and are planning the next prototype of the tutorial which will be evaluated by our established user base.

5 Conclusions

Existing speech recognition systems, although powerful, still present users with considerable usability problems. New users require a long learning curve before they become proficient in their use, and even long-term users have problems with recognition and error recovery. The systems cannot be used without recourse to keyboard, at least in the initial training stages. We are building a three-level tutorial system which will address these problems. An initial Hallway and Storefront evaluation indicated a number of interesting problems, but it showed that the main purpose of the Sampler system - to encourage new users to want to progress in the field of speech controlled computing - was a marked success.

6 References

Baker, J. 1989, "Large Vocabulary Speech Recognition", Speech Technology, April/May

CheepenC. 1990 The pragmatics of friendliness and user-friendliness, International Pragmatics Conference, Barcelona.

French, I, Halford, P, Hewitt, J, and Sapsford-Francis, J, 1994 "Developing Hands-Free Tutorials for Speech Controlled Systems in a Windows Environment", Computer Science Technical Report Number 195

Kay, P. 1991, "Speech Controlled Graphics on a Macintosh", in Independence through Technology, 7th Annual Conference, The BCS Disability Programme, Leeds; BCS, 45-50

Tough, C. 1990 "The Design of an Intelligent Transparent Speech Interface",

Vonk, R, 1990, "Prototyping", Prentice Hall

Zajicek M. & Hewitt J. 1990, "An Investigation into the use of error recovery dialogues in a user interface management system for speech recognition", in INTERACT'90 D. Diaper et al. (Editors), Elsevier Science Publishers B.V. (North Holland)