Proceedings of Intelligent Text and Image Handling - RIAO '91, April 1991, Barcelona

Technical Report No 119

James Monaghan - Speech & Language Technology Group
Christine Cheepen - Speech & Language Technology Group

April 1991

A System for Reducing Imprecision in Speech Interfaces to Generalised Text Input Devices

James Monaghan and Christine Cheepen
Speech & Language Technology Group
Hatfield Polytechnic, England

Abstract

This paper reports on ongoing research at Hatfield, where the Speech and Language Technology Group are developing a series of demonstrator systems in order to investigate human-computer interface design for systems where the input medium is speech. It discusses the ways in which ambiguities in input to speech systems can be avoided, how errors can be corrected, and describes the strategies which are being developed to deal with these problems at Hatfield and how these strategies are being evaluated.

Keywords: non-structured texts, speech interfaces, human factors

Introduction

Over recent years, much progress has been made in the mathematical and electronic optimisation of the technology underlying Automatic Speech Recognition (ASR) systems [1]. Progress in the parallel areas of the user interface and how users naturally talk in the relevant task-based contexts has only recently begun to make comparable strides. The present paper describes part of work of the Hatfield Polytechnic Speech and Language Technology Group's Intelligent Speech Driven Interface Project (ISDIP).

This project has produced a series of working applications, including a talking diary, a speech input word processor and a generalised speech interface. Some of the principles of the interface design underlying these applications will be described below, and the word processor and generalised speech interface will both be demonstrated. We will describe the context of this development and the underlying human factors and linguistic research. This takes account of the two most important strands in the study of natural dialogue between humans - the analysis of conversational form and the analysis of the underlying tasks which are represented and partially defined by the kind of conversation which takes place. The present paper will combine these two points of view and and draw on insights from a collaborative research project between Hatfield Polytechnic and

British Telecom on modelling dialogue between human and machine in the office context [2,3].

The greatest single contribution of Conversational Analysis to our understanding of the use of natural language in social contexts has been its demonstration of high level regularities in relatively informal aspects of language usage, such as telephone calls and conversations [4]. Work on openings and closings [5], on face work [6], interruptions [7] and repairs [8,9] has contributed much to the understanding of human-human interaction. Basically, it transpires that spoken interaction is not as irregular as we think and where apparent changes of plan take place and cause linguistic discontinuities they can be usually understood in terms of higher level strategies that require local fine tuning to fulfil. We will return to specific examples in dictation below.

In addition, work in task analysis in hierarchical goal-oriented environments has revealed a lot about hidden structures implemented beneath 'official' descriptions of social interaction [10]. As more and more human interaction is mediated by human-machine interaction, especially in contexts such as the office, it is important to apply tried and tested techniques of discourse and conversation analysis to this area. Not only do the two areas provide complementary insights but they both have structured analysis methods which have been established for other research purposes and so validate their use in the ASR context.

This paper discusses the use of such techniques to look at the restricted language within the office context with a view to isolating characteristic relationships of control, information processing and transfer, and strategies of text structuring and repair. Ways of applying these insights to an interface between a human user and modern electronic office systems will be discussed, and comparisons will be made with human-human interaction.

1. Data collection and analysis

The first stage of the Hatfield Polytechnic ISDIP project included a survey of human-human and human-machine interactions in working offices. The data is in the form of audio tapes of (1) people carrying out a range of office tasks (e.g. dictating a letter, making a telephone call, booking a diary appointment etc.), and (2) informal interviews with workers at various levels within the organisations being studied. The recorded material is being analysed with two primary aims in view - firstly the incorporation of natural conversational forms into the system, and secondly the close analysis of the specific forms of language used in particular domains within the office context, with a view to making the system 'aware' of what might reasonably be expected to occur within a particular task.

We wanted to identify the preferred linguistic forms which operate in a general way between people in the office context when engaged in particular tasks, in order to ensure that, where those tasks are performed by one person using a speech interface to a work station, the spoken dialogue between machine and user will, as far as possible, emulate the human-human forms. This is, of course, a very large and demanding area; we are dealing with it from three aspects - variability in user's commands, machine differentiation between commands and other forms of language, and repairs. During the course of the project, we will deal with the full range of office tasks (including spreadsheets, graphics packages etc.) but at the start we concentrated our work in this section of the project on the study of business letters - their content and the process of dictation.

1.1. In most conventional computer systems, the user has almost total responsibility for 'making sense' of the human/machine dialogue - that is, the system is inflexible, and the commands are only too often obscure to a novice user. We are aiming to allow the user to interact with the machine in as natural a way as possible, therefore we want to shift the main weight of responsibility for 'making sense' of the dialogue on to the machine. This does not simply mean using items from everyday language for the commands for each task, it means allowing the user to employ a *variety* of commands - in the case of dictating a letter, for example, the user might wish to use any one of: "dictation", "letter" "write a letter", "let's write a letter", "I want to write a letter". Work is currently in progress to implement this at the simple level of initiating a period of dictation.

1.2 Although letter dictation (followed by transcription) is one of the most frequent, run of the mill activities of the working office, the samples we have collected show that it is by no means a simple matter. The material which is dictated to a machine when the creator of the letter knows that another human being will process it before it becomes text is not simply a spoken record of the final contents of the letter, other kinds of language are also used - instructions to the secretary (e.g. punctuation of various kinds), personal, interactional remarks ("hi Lizzie", "there's another one snuffed it"), and asides ("if I can find it - I had it this morning").

To automate the process of dictation/transcription without placing possibly counter-productive restrictions upon the user, this mix of various language modes must still be available to a user interacting directly with a machine via the medium of speech. As the first step towards implementing this we have, as the example shows, analysed the dictation material in a broad way, by showing the differentiation between what is clearly intended as *text*, *instructions relating to the text* and *social, or interactional comment*. In computing terms, these three

kinds of talk can be viewed as *data*, *command*, and *comment*.

Key:

Italic	= interactional
Bold	= text of letter
Plain	= directive (initiation)
<u>Underline</u>	= directive (macro)
CAPITAL	= imperfect text (outline of memo)
[.]	= short pause
[-]	= long pause

hi Lizzie here's
some letters
for you . we'll start off with . er . .
Freeman and Baker paper file number 5106
that's er -
Mr and Mrs Brown .
write to the
Inspector of Taxes at **Brentford . -**
s er
J H and Mrs S A Brown . reference 655D. . . 54393 --- we regret to inform you that Mrs Brown died on the 12th of October 1988
---- *er .*
new paragraph ---
we note that we still await .
er
receipt of the . revised assessments for 1985 86 and 1986 87 . although we did receive a revised assessment for 1987 88 .
full stop new paragraph .
I don't normally do that do I - you're throwing me now -
we would like to point out that the personal allowance given is incorrect as this should be one thousand one hundred and sixty six pounds and not one thousand seven hundred and . fifty three pounds as stated - yours faithfully - .
can you do me
a memo on file number 1495
please
MR GRAY K R GRAY TELEPHONED ON . FRIDAY . 28TH OF OCTOBER . TO INDICATE THE BUILDING SOCIETY INTERESTS THAT ARE REQUIRED .
er if you could knock up
a memo from the details that I've scribbled down there
please - now we need to

write to
another one - there's somebody's snuffed it - erm ---- which will be .
file number 1008 Inspector of Taxes --- Finchley district **reference 276 stroke 40910 -**
you . er . .
C G Andrews - we . regret to inform you that our client died on the 14th of October -
new paragraph
we've been instructed by - the executor his son Mr A Andrews to
er
assist in . concluding taxation matters and hope to be able to write to you in detail . shortly .
new paragraph
we have in our possession . an income tax return for 1988 89 and you will no doubt require . further return . up to date therefore we should be pleased if you will let us have this in due course - yours faithfully .
write to .
the manager Abbey National Building Society . 50 Ballards Lane Finchley N32DP . C G Andrews -
erm .
and the address .
as you are no doubt aware we acted as accountants to the above who died on the 14th of October .
new paragraph
we have been instructed . by . his . executor Mr A (inaud.) returns to date .
new paragraph .
we have ascertained the details of interest on the - seven day account on the higher interest account with your society but are somewhat confused concerning . bond shares . which appeared to be in existence .
new paragraph -
according to notes we made last year ---------
if I can find them ----- I found them this morning ----------
there were some bond shares under reference JFY 1421535 which . were . cashed in . February 1987 . there appear to be other . bond shares under reference JFY 1543090 . on which interest was credited at March 1987 . and which therefore no doubt were not encashed until after 5th of April 1987 .
new paragraph
we shall be . grateful if you will let er know when these were cashed and indicate the amounts of interest . from 6th of April up to the date of encashment . yours . faithfully --

The next step in the research is to formalise the characteristics of each kind. Clearly, the area of *directive (initiation)* (e.g. punctuation) in the context of word processing is the simplest, in that it comprises only a small number of possible items (e.g. full stop, new paragraph etc). How to charactise the difference between *text* and *interactional*, however, is rather more complex, although, to the human hearer/reader, the distinction is immediately obvious. In order to build awareness of the distinction into our developing system we are looking first at differences in grammar and vocabulary - the importance of informal expressions like 'snuffed it' is clear in this context, but ultimately we also plan to build into the system a secondary check relating to pitch and intonation, as well as the use of filled and unfilled pauses (marked by -- etc., and by 'erm' etc respectively). This will be of use in cases where the grammar and vocabulary are not sufficient to make a clear distinction between two language modes.

2. Incorporating machine intelligence - a pilot study

2.1 Within the Hatfield ISDIP project we have conducted a pilot study on a small corpus of business letters - some 7,000 words - to test out the feasibility of incorporating knowledge of some of the recurring linguistic structures into our speech driven word processor, and to illustrate how such an implementation can improve the usability of the system. The following paragraphs explain why such an implementation is necessary for a speech input word processor, and give an account of how knowledge of some linguistic structures have been incorporated into our developing system.

The advantages of incorporating this kind of intelligence into a speech input word processor are twofold. Firstly, in cases where the input signal may be ambiguous (where the user's voice quality changes, or there is background noise) the User Interface Management System does not have to rely solely on the efficacy of the speech recognition device, but can also access the system's knowledge of what 'should come next'. This minimises the chances of misrecognitions and machine errors. Secondly, all speech recognition hardware is built in such a way as to constrain the size of the vocabularies used. In word processing, where a large number of vocabulary items is required, this means that the user must have access to a number of vocabulary sets. When the system has no expectation of linguistic structures the responsibility for remembering where the different vocabulary items are stored and for switching from one vocabulary to another rests with the user. In order to keep the vocabularies to a manageable size (and to avoid very similar items being stored together) it may be necessary to use as many as 20 or 30 different vocabulary sets in order to have access to a total vocabulary of 2,000 words. This is both an unacceptable load on the user, and a major factor in

slowing down the overall operation time. If the system has knowledge of expected linguistic structures, then the responsibility for vocabulary switching can be transferred from the user to the UIMS, thus making the system both faster and more usable.

2.2 The data for our pilot study consisted of 30 letters written from the research group to outside organisations and individuals. These texts were analysed using the Oxford Concordance Program, which gave us a total word count, word frequencies and concordances. The way this information was incorporated into our system can be illustrated by reference to two functional 'slots' common to all the letters - 'letter openers' (those items which occur after the salutation "Dear ...") and 'sentence openers' which occur elsewhere in the texts.

Throughout the corpus, there were 7 letter openers, which were also used as sentence openers together with 14 other items (including "yours sincerely" and "yours faithfully"), giving a total of 21 vocabulary items to account for every case of opening a sentence throughout the corpus.

These items were then used to build one vocabulary, which was automatically accessed by the system in two cases - the first when the word "Dear" had been recognised and followed by an item (or items) which ended with a comma, and the second when the item "full stop" had been recognised. The user was thus able to simply continue voice input of the text at these points without having to first summon the appropriate vocabulary. The potential for misrecognition was also greatly reduced, as the number of items in the vocabulary was very restricted, thus maximising the chances of a good match with the voice input. Such an implementation has, then, benefits for the system in terms of both usability and accuracy, which combine to give a considerable improvement in overall response time.

While the incorporation of this kind of knowledge into the system will drastically reduce the number of potential misrecognitions, errors of one kind or another will nevertheless occur, due to the limitations of speech recognition hardware and the variability of voice input, and any speech recognition system must provide for the repair of these errors. Our approach to this problem is dealt with in the following section.

3. Repairs

The importance of repairs and repair strategies has been investigated and illustrated in considerable detail within the field of Conversation Analysis. In

human-computer discourse, repairs are equally important, and, in the case of a human-computer speech interface (where speech may be the user's only form of access to the system), it is essential to provide for the correction of errors in machine output by the use of speech alone. When building a speech interface, it is important to build in repair procedures from the beginning, as misrecognitions will inevitably occur, and the user-machine dialogue must be able to overcome such communicative breakdown without the user resorting to extremes such as rebooting the system.

3.1 We have, to date, incorporated two different levels of repair procedure into the system to cope with misrecognitions. Where the word spoken is not unambiguously matched with a template within the machine, the system tells the user, offers a "best match" word, and asks for confirmation. The user is required to answer either yes or no. If the answer is no, the system offers a "next best match", and again the user is required to answer yes or no. If the answer is still no, the system prints nothing on the screen, and the user must input the word again.

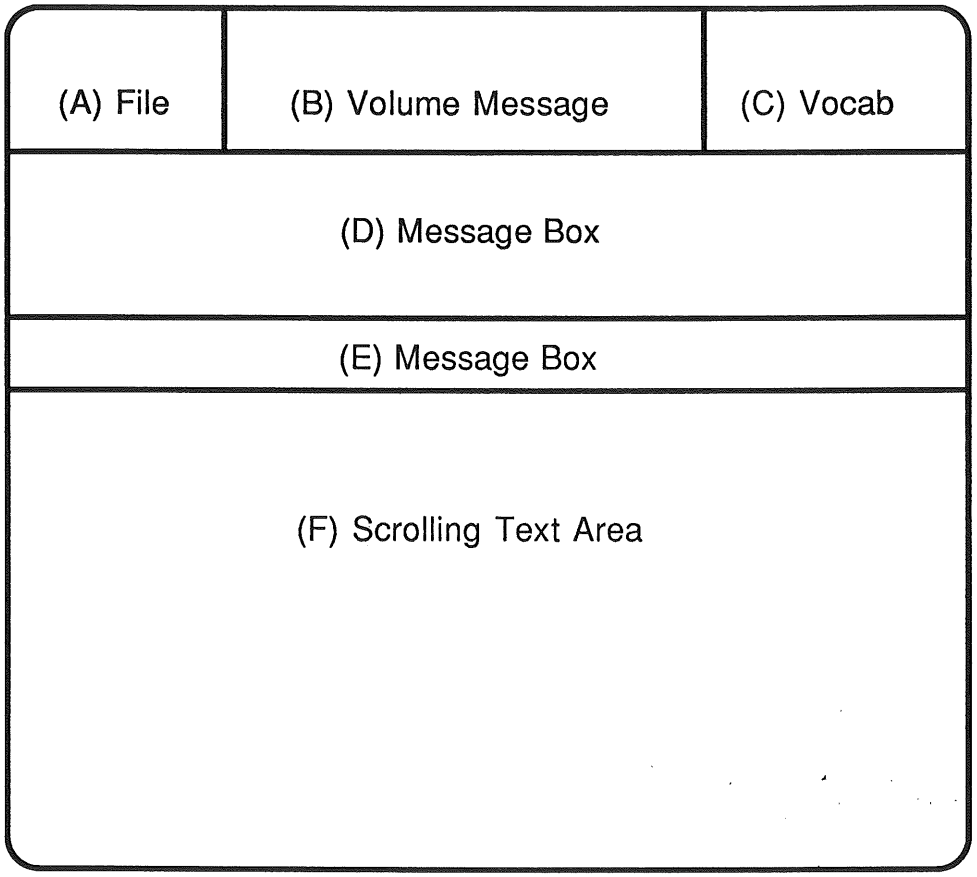
In practice, we have found this simple repair procedure, coupled with a very good recognition rate, to be extremely reliable, and multiple misrecognitions are very rare indeed. They must, however, be provided for, as the final system must be able to cope with a very wide range of human speech, including the 'inconsistent' speech of the voice impaired. In order to cope with this, we have built in a 'failsafe' mode, which goes into operation if the repair procedure is invoked three times in succession without a successful machine output. The system prints an error message on the screen together with a list of every item in the vocabulary currently in use, highlights each word in turn, and asks the user to whistle when the correct word is highlighted. When this has been done, the system asks the user if he/she would like to update their voice template by retraining the word. When this has been done, the system automatically returns to the word processing task in hand.

3.2 Although, in the examples we have used here, the person dictating the textual material is fluent, this is not always the case, and many users will wish to make (sometimes frequent) self corrections when producing the text. This is perhaps particularly the case where the output is a lengthy report, or a piece of academic work, the structure of which may be rather less constrained than a business letter. To provide for these cases, we have incorporated into our system a mechanism for the repair of user errors[1] - this is an *undo* command, where the last machine action can always be undone. Coupled with standard editing

[1] Our use of the term 'error' is intended to cover those cases where the user changes his/her mind about what item should come next in the text.

functions which enable the cursor to be moved around the screen and deletions, insertions and alterations to be performed.

4. The ISDIP word processor screen



A. The name of the file being edited.
B. Reserved for the messages 'Speak louder please' or 'Speak quieter please'.
C. The name of the vocabulary currently in use by the Recogniser.
D. The area for 'machine error' messages - i.e. cases where the system's recognition of the input is less than perfect, and it is offering 'best match' options.
E. The area for sub-dialogues between the user and the interface, and auxiliary error messages e.g. 'Please answer Yes or No only'.
F. The scrolling text area

5. Evaluations and ongoing work

5.1 We have evaluated the various repair strategies by a series of experiments with first time users [11], and the results are being used to improve the usability of the system following an iterative design approach.

The experiments were designed to ascertain whether first time users prefer an informative error recovery dialogue to a minimal one or none at all. The evaluations to date have run counter to our original expectations, and have led us to reconsider the interface design and the form of error recovery which should be offered. Our experiments indicate that there exists a point at which human expectation of the capabilities of a speech system and the conceptual model of the system which the user constructs cease to relate to standard models of human-computer interaction, and become closer to a model of natural conversation. Further trials are now needed in order to investigate the degree to which a 'natural' conversation can be emulated. We also intend to assess more experienced users in interaction with the system, since their perceptions of the usefulness of the various strategies may be different from those of the first time users already studied.

5.2 In parallel with the design-evaluation-redesign work on repair strategies, we are proceeding with the development of machine intelligence for incorporation into the Hatfield system. At present we are working on a large corpus of business letters donated to the project by a London solicitor. We are using the insights gained in the pilot study, and pursuing these in greater detail with this new corpus, in order to extend the capability of the system to predict the next user action and to automatically switch between vocabularies. Initial work has shown that, although the total corpus contains over 25,500 words, the total vocabulary is only 2,200 discrete items, which indicates that there will be a high degree of predictability for this restricted domain. Evaluation of this aspect of our work is built into our work plan, as we are using only 75% of the corpus for our analyses, with a view to using the results to predict the other 25%.

6. References

[1] Ainsworth, W. A. **Speech Recognition by Machine** (IEE Computing Series 12) Peter Peregrinus for the IEE, 1988

[2] Watkinson, N., J.Hewitt, C.Cheepen, J.Monaghan & J.Hobson 'The Office Context Usage Reference Model', Deliverable no. 3, Research Project A114929/P, Hatfield Polytechnic and British Telecom, 1990

[3] Hewitt, J., J.Monaghan, C.Cheepen & J.Hobson 'Recommendations', Deliverable no. 4, Research Project A114929/P, Hatfield Polytechnic and British Telecom, 1990

[4] Schegloff, E. 1968 'Sequencing in conversational openings', Laver, J. & S.Hutcheson **Communication in Face-to-Face Interaction: Selected Readings**, Harmondsworth, Penguin, 1972

[5] Schegloff, E. & H. Sacks 1973 'Opening up Closings', Turner, R. (ed) **Ethnomethodology**, Penguin, 1974

[6] Goffman, E. 1955 'On Face Work: an analysis of ritual elements in social interaction', Laver J. & S. Hutcheson (eds) **Communication in Face to Face Interaction**, Penguin, 1972

[7] French, P. & J.Local 'Turn-competitive incomings', **Journal of Pragmatics**, 6, 1982

[8] Cheepen, C. **The Predictability of Informal Conversation,** Pinter Publishers, 1988

[9] Cheepen, C. & J.Monaghan **Spoken English: A Practical Guide**, Pinter Publishers,1990

[10] Hewitt, J., J.Sapsford-Francis & J.Hobson 'An Application of Task Analysis to the Development of a Generic Office Model', **Human-Computer Interaction, Proceedings of Interact '90**, North-Holland, 1990

[11] Zajicek, M. & J.Hewitt 'An investigation into the use of error recovery dialogues in a user interface management system for speech recognition', **Human-Computer Interaction, Proceedings of Interact '90**, North-Holland, 1990