A ROBUST DIALOGUE CONTROL STRATEGY TO IMPROVE THE PERFORMANCE OF VOICE INTERACTIVE SYSTEMS

P C Millar, I R Cameron and D J Chaplin*

ABSTRACT

The performance of the current generation of automatic speech recognition hardware is generally poorer than that of human listeners. However, the effective recognition accuracy of voice input systems incorporating such devices can be increased by making careful use of interactive dialogues. The criteria for a good dialogue control strategy are discussed, and a report of a trial of a particular scheme is given. Results from this trial indicate that word misrecognition can be decreased by about 25% without compromising the fluency of the man/machine interaction.

INTRODUCTION

It is still the case that speaker-dependent word recognition is more accurate than speaker-independent word recognition, and that systems perform better when operating in isolated word mode than connected word mode. For many applications it is acceptable to constrain users to speak in isolated word format, but, in applications where there is a large user base, it can be extremely limiting if the word recognition is speaker dependent. The problem becomes even more acute if the system is to be used by the general public over the telephone network, because collection and storage of separate templates for every possible combination of voice types and channel distortions would be impracticable.

For these applications it is therefore necessary to use systems with a speaker-independent recognition capability. The system designers must however be aware that there will be a high incidence of word misrecognition. Hence a strategy for confirming or correcting responses will be needed to maximise the overall performance of such systems. It has been found that this can be done effectively by means of interactive confirmatory sub-dialogues that are initiated whenever the system detects that misrecognition is likely to have occurred.

CONFIRMATION STRATEGIES

The main requirements of sub-dialogues used for control are that they should make the overall interaction more efficient and they should not appear to be unnatural or tedious to the user.

Confirmation dialogue control could be as simple as making the machine ask for confirmation after every input. For example, if the machine thought the user said, "<X>"., the machine would say, "Did you say, <X>?

It would then expect a Yes/No reply before performing a task or continuing with the main dialogue. However, such a strategy would be laborious and time consuming for the user. It would, therefore, be inappropriate for the majority of applications.

Most word recognisers are, however, capable of returning confidence scores for at least the best two word matches, X1 and X2. With this

* The authors are at British Telecom Research Laboratories, Martlesham Heath, Ipswich, UK, IP5 7RE.
facility it is possible to make soft decisions and only invoke a confirmation routine when the scores are within a certain range of each other. The confirmation could involve asking for a repeat of the utterance, or using the "Did you say, <X1>?" response [1,2]. In practice the "Please repeat" response is only useful when the user's utterance is atypical due to noise masking or sloppy speech. In most other situations the repeated utterance would be just as likely to be confused as the first. Similarly, a "Did you say, <X1>" request is only useful if <X1> was in fact said. If it were not then supplementary questions have to be asked. These supplementaries could involve queries about the second or even the third etc. best choices [3,4]. In practice, however, the third choice word is rarely correct, and the extended list querying option is often found to be frustrating for the users, because they usually want to say their word again if the first query proves inconclusive.

A preferred strategy is therefore to ask, "Did you say, <X1>?" and if the reply is, "No", the machine should continue, "I see, then did you say, <X2>?" If this second choice is also incorrect the machine should go on to ask the user to repeat their utterance. This particular strategy is similar to the confirmation procedure that normally takes place during a human-human interaction. The dialogue therefore appears natural and concise.

**DIALOGUE PERFORMANCE TESTS**

The usefulness of this control dialogue was tested on a voice interactive system that is intended for operation over the telephone network by members of the public. The system therefore has to be simple to use and must be tolerant to a wide range of voice types and to distortions introduced by a variety of different transmission channels. The target service used in these tests was a telephone banking application.

Twenty eight subjects (15 male, 13 female) took part. The tests were carried out over a telephone connection to the voice response system. The subjects were asked to perform a set of ten multi-response tasks, which exercised the complete vocabulary of the system. None of the subjects had been trained on the system so each was given an explanatory sheet describing how to use the system and its limitations.

The words in the vocabulary of the recognition unit were:

<table>
<thead>
<tr>
<th>statement</th>
<th>transfer</th>
<th>Visa</th>
<th>twenty</th>
<th>full</th>
</tr>
</thead>
<tbody>
<tr>
<td>chequebook</td>
<td>savings</td>
<td>one</td>
<td>thirty</td>
<td>stop</td>
</tr>
<tr>
<td>mini-statement</td>
<td>current</td>
<td>two</td>
<td>forty</td>
<td>yes</td>
</tr>
<tr>
<td>balance</td>
<td>Access</td>
<td>ten</td>
<td>fifty</td>
<td>no</td>
</tr>
</tbody>
</table>

Part of a typical dialogue is given below. Note that the confirmatory, error-correcting routine is only invoked when the machine is uncertain about the user's response, in this case during the chequebook request.

**Machine**, "What service do you require?" **Subject**, "Balance."
**Machine**, "The balance of your current account is 203 pounds and 32 pence. Do you require another service?" **Subject**, "Yes."
**Machine**, "What service do you require?" **Subject**, "Chequebook." (Word masked by noise.) **Subject**, "No."
**Machine**, "Did you say, transfer?" **Subject**, "Yes."
Machine, "Confirmed. A new chequebook will be sent to you today. Do you require another service?"... and so on.

It is good practice for systems of this sort to obtain a positive confirmation before instigating any consequential actions. So, had the soft decision confirmation not been triggered for the chequebook request above, the machine would have continued with, "You have requested a chequebook, is that correct?", and this would have necessitated a further reply from the user. It can therefore be seen that the confirmatory sub-dialogue ought not increase the overall time of the interaction. In fact it can effectively shorten it by correcting errors immediately and obviating the need to invoke the service confirmation routines.

For a fluent dialogue, it is obviously important to establish a sensible score difference threshold for triggering the soft confirmation routine. If it is set too low then words will be misrecognised frequently, but if the threshold is too high then most of the confirmation requests will be merely substantiating the machine's first choice guess, and that becomes tiresome for the user. The plot in Fig. 1 shows the ratio between the useful confirmation requests (2nd choice correct), and the non-useful requests (ie 1st choice correct), as the score difference between the 1st and 2nd choices increases. For our system, we arbitrarily chose a score difference of 7 to be the confirmatory threshold, and this resulted in about 4.3 times as many non-useful confirmations as useful ones. This was considered to be acceptable from the point of view of fluency of the dialogue.

![Fig.1 Plot of score difference between 1st and 2nd choice words against the ratio of 1st choice correct and 2nd choice correct.](image)

With the threshold set at this level the performance of the system was analysed for the tests undertaken by the panel of subjects.

RESULTS

The system's responses have been categorised as follows:

1. The word spoken was recognised correctly first time
2. The user was asked to confirm the system's 1st choice word (unnecessary confirmation)
3. The 1st choice word was incorrect but the 2nd choice was correct (useful confirmation)
4. Both the 1st and 2nd choice words were incorrect (non-confirmation)
5. The number of times where NO or YES were confused (YES/NO mismatch)
6. The word spoken was incorrectly interpreted and not queried (misrecognition)
7. No matching word was obtained and the user was asked to repeat their utterance
(8) The user was thrown off the system because the machine failed to recognise a word on 3 consecutive occasions (fail).

A total of 1528 words were spoken by the trialists as they attempted to perform a total of 280 tasks. Their responses were as follows.

<table>
<thead>
<tr>
<th>Category</th>
<th>Number of words</th>
<th>Percentage of total</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) (Correct 1st time)</td>
<td>1291</td>
<td>84.49%</td>
</tr>
<tr>
<td>(2) (Unnecessary confirmations)</td>
<td>125</td>
<td>8.18%</td>
</tr>
<tr>
<td>(3) (Useful confirmations)</td>
<td>29</td>
<td>1.90%</td>
</tr>
<tr>
<td>(4) (Non-useful confirmation)</td>
<td>10</td>
<td>0.65%</td>
</tr>
<tr>
<td>(5) (YES/NO mismatch)</td>
<td>5</td>
<td>0.33%</td>
</tr>
<tr>
<td>(6) (Misrecognition)</td>
<td>9</td>
<td>0.59%</td>
</tr>
<tr>
<td>(7) (No match)</td>
<td>59</td>
<td>3.86%</td>
</tr>
<tr>
<td>(8) (Fail)</td>
<td>6</td>
<td>0.39%</td>
</tr>
</tbody>
</table>

This indicates that the confirmatory dialogue has increased the word recognition accuracy of the machine from 92.7% to 94.6%. More detailed analysis of the results showed that twenty four out of the twenty eight subjects (86%) could perform all the tasks satisfactorily. Of the four that had difficulty, three had difficulty with one response only, and the remaining subject had trouble with two responses. There was one failure due to a YES/NO mismatch, and one occasion when the dialogue became confused, but this did not result in a service being instigated in error. Thus, from a total of 280 tasks that were attempted, 274 were performed successfully (97.9%). Additional tests that were performed without the confirmatory sub-dialogue resulted in a lower task success rate (92.3%), and required the users to speak about 5% more words. Thus our claim that the control dialogue improves accuracy but does not lengthen the overall time of the complete interaction also appears to have been substantiated.

CONCLUSION

Careful use of confirmation loops and soft decisions during man-machine dialogues can improve the performance of the system. The interactions can become more natural for the users, and false word recognition can be decreased, in this particular case by around 25%. Similarly, multi-response task failure rates can be substantially reduced, for this target application by around 70%.

ACKNOWLEDGEMENT

Acknowledgement is given to the Director of British Telecom Research Laboratories for permission to publish this work.

REFERENCES