Proceedings of the Institute of Acoustics

RESULTS OF AN EXERCISE TO COLLECT 'GENUINE' SPOKEN ENQUIRIES USING WOZ TECHNIQUES

R K Moore and S R Browning

Speech Research Unit, Defence Research Agency, St. Andrews Road, Malvern, Worcs., WR14 3PS

1. INTRODUCTION

Many laboratories have now used the so–called 'Wizard of Oz' (WOZ) technique for eliciting spontaneous spoken human–machine dialogue in order (i) to study the resulting speech and (ii) to evaluate the necessary speech technology and natural language processing systems, Fraser [1]. The technique involves a human expert (the wizard) taking the place of all or part of an automated system, and the user interacts with the simulated system, believing that it is a machine. The WOZ protocol enables users' behaviour to be studied under conditions which are not constrained by current technological (or theoretical) limitations.

However, many such exercises involve 'volunteer' users whose behaviour is prescribed by a pre–prepared scenario, Zue [2], thereby removing one potentially crucial aspect of human–machine interaction, namely any behavioural events which are unique to *genuine* (i.e. motivated) and un–preprepared transactions, Spitz [3].

This paper presents the results of an exercise to collect unscripted speech data using the WOZ technique by the provision of a *genuine* voice–based telephone enquiry service. A direct comparison is made with the equivalent non–WOZ service.

2. THE TASK DOMAIN

The enquiry service was configured around 'AUTOROUTE' – a commercially available route planning software package, Nextbase [4]. The package runs on a PC and contains map and gazetteer information covering the majority of the United Kingdom. Its main feature is its ability to find the shortest and/or quickest routes between two locations in accordance with a range of specifiable variables such as preferences for certain classes of roads and driving speeds. Alternative routes can also be found.

For the purposes of this exercise, the route planning package was configured to show no particular preference for road type and the road speeds were set at the national speed limits.

3. THE WIZARD

Clearly the behaviour of the wizard strongly influences the nature of the resulting corpus, and constraints (such as restricting the vocabulary) can be placed on a wizard in a variety of ways, Moore [5]. For example, the accepted vocabulary might be restricted, or the complexity of the language that can be understood might be limited in some way.

However, in order not to restrict or influence users' behaviour, it was decided that very few restrictions should be placed on the wizard apart from the use of a stock opening phrase and the provision of a few standard reply templates simply in order to reduce the wizard's work load. Also only questions to do with the chosen task domain would be accepted; all other enquiries about the system itself, or who was running it should be met with a standard response. This was intended to elicit the most natural behaviour possible from the users, while keeping them within the task domain.

In general, the design for the wizard's behaviour was based on information derived from the procedures employed by a commercial company who already provide a route planning service over the telephone (in this case the enquiries being made by tone dialling) based on the same software package.

Since the application was to be telephone–based, all interaction with the user (input and output) had to be conducted using speech. Therefore, in order to create the illusion that the operator was a machine, it was important that the output to the user should sound 'mechanical' i.e. like synthesised speech. However, it was considered that the quality of contemporary speech synthesis was not high enough for the purposes of this exercise since it was anticipated that the quality of the wizard's voice would affect the user's perception of the system's capabilities; high voice quality being likely to suggest a system of high capabilities while a low voice quality would not only imply a system of poor capabilities but might lead to excessive confirmatory dialogue if the user had difficulty understanding the response, Moore [5]. It was therefore decided that the characteristics of the operator's natural voice should be altered electronically. A suitable device was constructed which was based on the principle of changing the pitch of the operator's voice and then combining the natural and altered signals to produce a highly synchronised duet effect. It was found that this provided a voice which was both unnatural (indeed 'robotic') and yet fully intelligible, Taylor [6], and that when the operators heard their own voices changed in this way they found themselves speaking in almost a monotone, hence increasing the unnatural effect.

In addition, the wizard was instructed to make every attempt to remove all distinctly human characteristics from their speech such as false starts and stutters, and great care was taken to ensure that breath noise and key clicks were not audible to the caller. Also, the wizard was instructed to pause for a couple of seconds before responding to the users' queries, not only in order to give the impression that their input was being 'processed', but also to allow the operator to plan the response, so as to minimise the false starts or stumbling that are typical of human speech.

4. THE EXPERIMENTAL CONFIGURATION

As mentioned above, as well as implementing a genuine telephone–based enquiry service using WOZ techniques, it was also decided to compare wizard–type transactions with normal human–human interaction for the same task. Hence the experimental set–up was configured to operate with two incoming telephone lines – one assigned to the normal human operator and one assigned to the wizard. Appropriate equipment was installed to provide automatic detection of incoming calls and initiation of recording and digitisation. All recording sessions took place with the operator in a sound proof booth.

A modified Ensigma DSP32C telephony board, Ensigma [7], was used to detect an incoming call and took the line off hook (i.e answered it). A computer program controlled the telephony board and the recording of both sides of the conversation to hard disk, as well as doing some housekeeping, such as updating filenames. When ringing was detected the computer beeped to let the operator know. The operator then instructed the program to start recording to disk, and also started a DAT recorder which was employed to record the interaction as a back–up. The operator spoke into a Shure SM10 close–talking headset microphone.

During the recording sessions the relevant line was connected to the computer, while the other was attached to an answering machine, on which was a message directing callers to the other number. When the service was not "on air" both lines were connected to answering machines, again with messages telling callers when the service would be available.

As the interaction proceeded the operator typed in the user's query into Autoroute running on a PC, and when the program had produced its answer the operator summarised it to the user, offered to send them the table of instructions, and took a note of their name and location. This was useful as it was intended to send out a questionnaire at a late date to ask users what they thought about the service. Callers were also given the option of having the route information read out to them during the call, but this was not encouraged, as it is a lengthy process (and placed great stress on the wizard). At the end of the interaction the operator instructed the computer to stop recording, and also stopped the DAT recording. The telephony board put the telephone back on–hook, ready for the next call. A manual log was kept, summarising the content of the calls.

In order for there to be minimal differences between the operator's behaviour in both the human–human and human–wizard conditions, the same operator was used in each case. As a consequence, the only difference between the two conditions was that the wizard's natural voice was modified using the 'voice disguise' unit.

5. INSTRUCTIONS TO THE USERS

It was decided that the population of potential users should be staff from within the Defence Research Agency. An initial pilot phase was carried out using only staff from the Malvern site, followed by a main phase which expanded the operation to cover the entire Agency.

For the pilot phase, a poster advertising the service was circulated for display on site noticeboards and an electronic advertisement was placed in the central computing facility. For the main phase, the poster was replaced by a small leaflet which was placed on the desk of each DRA employee (see figure 1).

Since the emphasis of the exercise was to collect data from genuine enquiries, it was necessary to take great care to ensure that no–one knew the true nature of the experiment or who was running it. It was also necessary to consider carefully the wording of the advertisements since it is known that it might have the wording of any instructions in such simulation experiments has a strong influence on how users choose to express their queries, and indeed on what kinds of questions they might try to ask; it was important not to put words into the users' mouths. As a consequence, the advertisements made no mention of either the Speech Research Unit or of the difference between the two available telephone numbers, nor did they specify that the service was experimental, computer–based or automatic in any way.

COLLECTING 'GENUINE' SPOKEN ENQUIRIES USING WOZ TECHNIQUES

Figure 1. The advertisement for the route planning service.

On receipt of a call, the operator (in human or wizard mode) always used the following introductory announcement:– "Welcome to the route planning service – how can I help you?". This message assured users that they had dialled the right number and gave a very broad idea of what the service was about. No mention was made of whether the service was automatic or not; users were left to infer this for themselves on hearing the introductory message.

In the wizard condition feedback was usually given to the user to confirm that the system had 'understood' the user's request, and to give information about what the system was doing (e.g. "Please wait while the route between Malvern and Edinburgh is being calculated"), to explain some of the (sometimes quite long) silences that occurred while the wizard was typing the query into AUTOROUTE. Similar feedback was also given at times in the human mode, but tended to be less frequent (as confirmation that the operator had understood was needed less often) and more informal (e.g. "Ok it's just doing that for you now").

In the wizard mode, enquiries about the service itself were met with a standard response: "This service can help you find a route between two places in the United Kingdom", and further questions were countered by "I'm sorry but for commercial reasons I can not give you any more information".

6. THE RECORDING EXERCISES

The pilot recording phase ran for a three week period in November 1991, and is described in more detail in Morris [8]. The posters advertising the service were distributed about a week beforehand. The main phase, described in Browning [9] drew on the lessons from the pilot study, and recordings were made during the first two weeks of April 1992 – this period being specifically chosen as it involved the weeks leading up to Easter holiday, when it was expected that many people would be planning journeys. Although the set up for the experiments was fundamentally the same for both phases, the operators were different; the pilot phase was run by a temporary attachment to the SRU, while the main phase was conducted by a permanent member of staff.

The service was made available on each line for alternate half day sessions from 10 a.m. to 12 a.m. and 2 p.m. to 4 p.m. Whilst one number was on–line the other was connected to an answering machine which requested the caller to try the alternative number. The time was split equally between human and wizard conditions.

The pilot phase recordings were transcribed by the departmental PA and the main–phase recordings were transcribed at SRI International in Cambridge.

8. RESULTS

During the pilot phase (13th of November 1991 to 5th of December 1991) the service received a total of twenty–two calls, and these are summarised in Table 1. Of these, sixteen came to the wizard and six to the human. All of the calls to the human operator were genuine enquiries, either asking for a route, or enquiring about the service. Three of the wizard's callers hung up without speaking, and another one gave up halfway through the conversation. It is not possible to decide how many of the rest of the calls to the wizard were genuine, or whether some callers were just experimenting with the system.

	Total calls	Route enquiries	Service info	Hang –ups	Gave up	Invalid calls
Wizard Operator	16	12	0	3	1	?
Human Operator	6	4	2	0	0	0

Table 1. Summary of calls during the pilot phase.

COLLECTING 'GENUINE' SPOKEN ENQUIRIES USING WOZ TECHNIQUES

A preliminary analysis of the pilot–phase transcripts produced a variety of interesting statistics on caller behaviour. The average length of a wizard operated call was two minutes, while those to the human were approximately three minutes long. Also, it was found that, on average, there were significantly fewer words spoken by the caller in each turn of the human–wizard condition than in the human–human condition and, although the rate of "uhms" and "errs" was about the same in both conditions, callers seemed to be more polite to the machine than to the human operator! In some calls there was considerable background office noise (some callers appeared to be using loud–speaking telephones) and callers occasionally chuckled to themselves or made asides to other people in their vicinity (including statements along the lines of "Hey, I'm talking to a machine" and constant references to 'it') – although this confirmed that the callers were convinced by the wizard's voice it also indicated that they believed that the system automatically knew when it was being addressed! Some callers interrupted the wizard, and one mimicked the robotic style of the wizard's voice. Moore [10] presents a more detailed analysis of the pilot phase data.

In the main phase the total number of calls received during the two week period was forty–four, of which thirty–one were taken by the wizard, and thirteen by the human. These are summarised in Table 2. Again, around a quarter of the calls to the wizard hung up without saying anything; about the same ratio as did so in the pilot phase. It was not possible to determine whether some of these people simply did not wait for the phone to be answered, or whether they were scared off by the artificial voice. In addition, two of the calls to the wizard were invalid; one because the caller recognised the wizard; the other was from someone who knew about the experiment. One caller treated the wizard like an answering machine and left a message! It is reasonably certain that all other calls were genuine.

	Total calls	Route enquiries	Service info	Hang –ups	Gave up	Invalid calls
Wizard Operator	31	14	7	8	0	2
Human Operator	13	10	3	0	0	0

Table 2. Summary of calls during the main phase.

The average length of calls to the wizard in the main phase was between two and three minutes, while to the human it was approximately four minutes. Again, the calls to the wizard appear to be more business–like, and callers were less inclined to digress from the subject. One caller to the human discussed the merits of the routes given. Contrary to what was found in the pilot phase, many callers asked the wizard about the service.

A detailed analysis was conducted on the main–phase transcripts and the results are summarised in figure 3. The bulk of the analysis was related to the callers' 'turn' behaviour, where a turn was judged to include all utterances intended to be part of the transaction (i.e. not asides or remarks to someone else) which were separated from the user's previous turn by the operator speaking or by something significant happening (mainly when the operator has two consecutive turns, between which the caller does not say anything, but AUTOROUTE produced some information which the wizard would then pass on). Apart from the results for "uhms/errs" and false–starts, the statistics of the transactions are very similar to those found in the pilot phase. As before, the wizard–based interactions were found to be significantly shorter that the human–based ones, although this was not due to simple one or two–word turns by the user (as might have been thought).

In the pilot phase, roughly equal proportions of "uhms/errs" and false–starts were discovered. However, in the main phase, such behaviour was much more frequent in the human–based transactions. This provides some indication of the relatively higher degree of user 'preparedness' in the wizard–based condition.

	WIZARD	HUMAN
Average no. of turns/call	7.5	26.5
Average no. of words/call	30.7	185.6
Average no. of words/turn	5.0	7.0
Turns with "uhms" & false starts	17.4%	39.5%
Turns with terms of politeness	31.0%	9.6%
One–word turns	23.2%	25.3%
One/two–word turns	48.4%	39.2%

Figure 3. Analysis of caller behaviour in the main–phase recordings.

In both experiments there was considerable evidence (mainly from background remarks) that users were convinced they were talking to a computer. It was also noticeable that, although the human–wizard dialogues were all concerned with planning particular routes, most of the human–human dialogues were about the nature of the service itself. In other words, the users who dealt with the wizard seemed to assume that such a system would not be able to provide explanations about what it could and couldn't do – and so they didn't ask.

9. CONCLUSIONS

This paper has described an exercise to collect unscripted speech data using the Wizard of Oz technique to provide a *genuine* telephone–based route planning service. Although only a limited quantity of data has been collected, several valuable insights into the nature of future speech–based human–machine interaction have been obtained. In particular, various practical problems have been highlighted such as the need to handle significant background noises, spoken asides by the user and interruptions. More importantly, however, it confirms that *genuine* spoken human–machine interaction is somewhat simpler and more (short–term) goal directed than corresponding human–human dialogue.

10. REFERENCES

[1] N M FRASER & G N GILBERT, 'Simulating speech systems', Computer Speech and Language, Vol.5, No.1, pp 81–99, January 1991.

[2] V ZUE, N DALY, J GLASS, H LEUNG, M PHILLIPS, J POLFRONI, S SENEFF & M SOCLOF, 'The collection and preliminary analysis of a spontaneous speech database', Proc. DARPA Speech and Natural Language Workshop, pp 126–134, Harwichport, MA, 15–18 October 1989.

[3] J SPITZ, 'Collection and analysis of data from real users: implications for speech recognition / understanding systems', Proc. DARPA Speech and Natural Language Workshop, pp 164–169, Pacific–Grove, CA, 19–22 February 1991.

[4] NEXTBASE LTD., Autoroute Plus User Guide, 1991.

[5] R K MOORE, M J TOMLINSON & A MORRIS, 'Whither the wizard?', Proc. ESCA workshop on the Structure of Multimodal Dialogue, Maratea, Italy, 16–20 September 1991.

[6] M M TAYLOR, *private communication*, September 1991.

[7] ENSIGMA LTD., DSP32C Telephony Board User Guide, Version 1.0, 1990.

[8] A MORRIS, 'A Wizard of Oz technique for capturing utterances of unscripted speech', DRA Malvern, IS2 Research Note No. 181, 1992.

[9] S R BROWNING, 'Collecting a corpus of unscripted speech using a Wizard of Oz technique', DRA Memorandum No. 4675, 1992.

[10] R K MOORE & A MORRIS, 'Experiences collecting genuine spoken enquiries using WOZ techniques', Proc. DARPA Speech and Natural Language Workshop, Harriman, NY, February 1992.

[11] J POLIFRONI, S SENEFF & V ZUE, 'Collection of spontaneous speech for the ATIS domain and comparative analyses of data collected at MIT and TI', Proc. DARPA Speech and Natural Language Workshop, pp 360–365, Pacific–Grove, CA, 19–22 February 1991.

(C) British Crown Copyright, 1992