Behaviour & Information Technology

ISSN: 0144-929X (Print) 1362-3001 (Online) Journal homepage: https://www.tandfonline.com/loi/tbit20

An experimental evaluation of preferences for data entry method in automated telephone services

J.C. Foster , F.R. McInnes , M.A. Jack , S. Love , R.T. Dutton , I.A. Nairn & L.S. White

To cite this article: J.C. Foster , F.R. McInnes , M.A. Jack , S. Love , R.T. Dutton , I.A. Nairn & L.S. White (1998) An experimental evaluation of preferences for data entry method in automated telephone services, Behaviour & Information Technology, 17:2, 82-92, DOI: 10.1080/014492998119571

To link to this article: https://doi.org/10.1080/014492998119571

Published online: 08 Nov 2010.

Submit your article to this journal

Article views: 42

View related articles

Citing articles: 1 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tbit20

An experimental evaluation of preferences for data entry method in automated telephone services

J.C. FOSTER, F.R. MCINNES, M.A. JACK, S. LOVE, R.T. DUTTON, I.A. NAIRN and L.S. WHITE

The Centre for Communication Interface Research, Department of Electrical Engineering, The University of Edinburgh, 80 South Bridge, Edinburgh EH1 1HN, Scotland

Abstract. This paper reports an experiment to investigate users' preferences amongst three modes of data entry in an automated home shopping service: *DTMF* input on the telephone keypad, and *isolated word (IW)* and *connected word (CW)* speech input. Preferences were measured both by means of attitude questionnaires and by giving participants an explicit choice among the three versions of the service once they had experienced them all. Users' attitudes to the service with a given mode of data entry were found to vary according to their cognitive skills (verbal and spatial abilities) and according to whether they had previously experienced a different data entry mode. Overall, DTMF and CW were rated similarly, and were strongly preferred to IW. Implications of these findings for the implementation of telephone-based services are discussed.

Introduction

The choice of data entry mode for an automated telephone service often lies between *DTMF* (dual-tone multi-frequency) input using the telephone keypad and one of two forms of speech input, viz. *connected word* (CW), where the user says a string of words following the service prompt, usually without pauses between words, and *isolated word* (IW), where the user says a single word following a service prompt.

Usability and human factors considerations dictate that users' preferences as to the mode of input should play an important part in the implementation of new services. This paper presents results from an experiment aimed at evaluating user preferences among these three input modes.

User preferences for input mode will depend upon a complex of factors such as perceived ease of use for entering data, the accuracy of response by the service and the perceived naturalness of the input mode as a medium of communication with an automated service.

The perceived ease of use of the telephone keypad as a data input mode will depend heavily upon the nature of the input vocabularies required by the service. Where the service requires only digit entry or the use of a very small vocabulary such as 'yes'/'no', the mapping between vocabulary items and keypresses is direct and simple. In such cases DTMF may well be perceived as easy and therefore be the preferred mode of input for automated telephone services. Where the task requires larger or more complex vocabularies (e.g. money amounts or alphanumeric vocabularies), the mapping from user goal to user action may be so complex as to make DTMF difficult to use and therefore not the preferred mode of data entry.

Furthermore, there may be a dislike of DTMF simply because it involves keying responses on the telephone keypad. This may arise from a lack of familiarity with keypads or a failure to perceive the telephone keypad as a means for making responses during a call.

CW has been claimed to be the most 'natural' input mode because it most nearly approximates to typical human-human discourse (Lea 1980, Brooks 1989). Several features of CW input to computers militate against this view. Firstly, the technology for recognizing connected speech, especially over the telephone network, is very error-prone (Waterworth 1984, Wilpon *et al.* 1990). Hence much of the user's interaction time with the service may be spent in error recovery, with the service asking for repetitions and requiring confirmations of inputs, perhaps associated with irritating reminders about how and when words should be spoken. None of this activity takes the user closer to achieving the user's task or service goals. Also it is easy to exaggerate the 'naturalness' of CW as a mode of communication with machines. Speaking a sequence of

0144-929X/98 $12.00 © 1998 Taylor & Francis Ltd.

words following a service prompt is far removed from the fluent rhythms of everyday human – human conversation (Karis and Dobroth 1991). People can, of course, learn to adapt to this method of entering data, but it remains an artificial and learned, rather than a natural, method of communication. The question remains whether this mode, once learned, is perceived as preferable to DTMF or IW.

IW is very definitely not a natural method of communication. It is therefore an example of using a 'natural' medium of communication (speech) in an artificial and highly constrained way (single, isolated words). The advantage IW has over CW as far as communication with computers is concerned is that recognition error rates are much lower than for CW, at least when measured per utterance rather than per word (Wilpon *et al*. 1990). Users are thus less likely to find themselves drawn into prolonged error recovery routines.

Little empirical work has been carried out into the relative preferences for data entry mode among users of automated telephone services. Many previous studies have concentrated exclusively on DTMF-based systems (Roberts and Engelbeck 1989, Martin *et al.* 1990, Schumacher 1992), while others have been concerned only with systems using speech input (Jack *et al.* 1992, Rosenbeck and Baungaard 1992). Some researchers have reported systems incorporating both speech and DTMF for different parts of the task, or using DTMF if available with speech as a fallback, but without making a direct comparison between the modalities (Levas 1988, Hammond 1990, Bossemeyer and Schwab 1991). However, one paper which did attempt to address the issue of relative preferences was Fay (1993).

Following a series of five experiments into preferences and attitudes towards DTMF and speech as input modes, Fay (1993) hypothesized that users' preferences depended on the design of the interface into which the DTMF or speech input was embedded: most users preferred the input mode, whether DTMF or speech, for which the interface was originally designed (and therefore presumably optimized) rather than the mode to which it was subsequently adapted. No differences in preference were found between samples of college students and of people recruited from a shopping mall, but correlations were found with subjects' attitudes to technology: those who were more positive towards telephony tended to prefer DTMF, while those who were more positive towards novel consumer electronics tended to prefer speech.

The experiences reported by Hornstein (1994) confirm that the preference between DTMF and speech input depends on details of interface design. In the first version of a banking service offering a choice between DTMF and isolated word input, users preferred DTMF, but when some problems with the interface were corrected the preference switched to speech.

The experiment reported in this paper investigated users' preferences amongst DTMF, isolated word and connected word modes for entry of digit sequences in an automated home shopping service. Each subject in the experiment used all three versions of the service, completing an attitude questionnaire after each, and then made an explicit choice amongst them for a fourth use of the service. Measurements of subjects' performance in the task with each version of the system were also taken. The subjects' personality traits and cognitive skills were measured using psychometric tests, and the relationships between these user characteristics and the attitude and performance measures were explored.

The immediate background to this experiment was a previous (unpublished) experiment conducted over the telephone network to compare users' attitudes to versions of the home shopping service with IW and CW speech input. This yielded the unexpected result that the two groups of subjects who experienced the different input modes had almost identical attitudes to the service, even though the recognition accuracy for both was set at 100% so that the higher rate of errors with CW input did not occur: under these conditions it had been expected that CW would be rated better than IW. To investigate further, it was decided to carry out another experiment, as reported here, with the same service and the same 100% recognition accuracy, in which all subjects would use both versions and then make an explicit choice between them. The opportunity was taken at the same time to introduce DTMF as a third mode of input and compare it with both speech input modes.

2. Experimental method

2.1. *Simulation methodology*

The experiment reported in this paper was part of an ongoing series of experiments using a real-time Wizard of Oz (WOZ) methodology designed to permit the investigation of users' attitudes towards simulated telephone services and the evaluation of the perceived usability of such services (Foster *et al.* 1992, 1993, Jack *et al.* 1992). During the experiments, specific features of the human – computer interface are manipulated and the dependent variable of user attitude is measured using questionnaires.

Among the features of the interface which can be independently modified in the WOZ experiments are a simulated speech recognizer and a dialogue scheme. One of the new features of this work is that it is based on the parametric simulation of existing speech recognition technologies. This allows experiments to be carried out both with current recognition performance and with recognition performance extrapolated beyond that currently achievable. In this way, it is possible to address, among other issues, the shape of the usability function for automated telephone interfaces for different levels of recognition performance (Jack *et al.* 1992). Modifying the dialogue scheme allows important user interface and human factors issues to be addressed such as the impact of voice quality, the degree to which conventional tone prompts (used in addition to spoken prompts) influence the progress of the dialogue, and the impact of dialogue structure and prompting strategies on users' responses to the service.

WOZ experiments (Fraser and Gilbert 1991) involve the use of a hidden operator who, unknown to the experimental subject, simulates some aspect of the performance of a computer. In the case of the present experiments, contact between subject and computer is over the telephone line. Consequently, it is a relatively easy matter to ensure that subjects are not aware that a human being plays any part in the running of the automated service.

The WOZ experimental configuration which includes the user/subject, the WOZ operator and the speech interface software is shown in figure 1. In experiments, the software control module handles the initiation of contact over the telephone network, the delivery of the dialogue prompts to the subject, the registering of keystrokes by the WOZ operator made in response to the spoken input from the subject, the on-line generation of recognition errors when these are required and the recording of all data on keystrokes and timings. The WOZ software runs on an IBM-compatible PC with a plug-in circuit card providing connection to the telephone line.

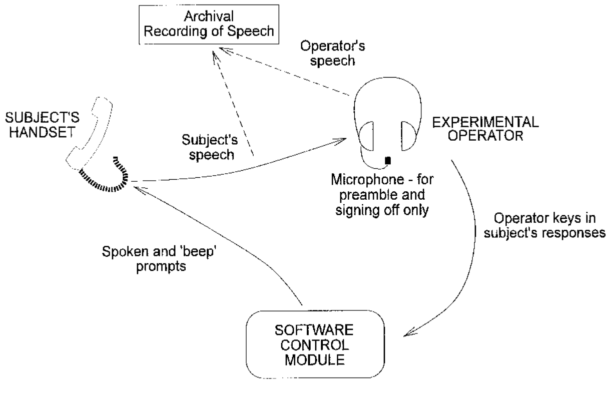
Archival Recording of Speech
Operator's speech
SUBJECT'S HANDSET
Subject's speech
EXPERIMENTAL OPERATOR
Microphone - for preamble and signing off only
Operator keys in subject's responses
Spoken and 'beep' prompts
SOFTWARE CONTROL MODULE

Figure 1. The Wizard of Oz experimental configuration.

This particular WOZ setup differs from many previous WOZ configurations reported in the literature (Fraser and Gilbert 1991) in the degree of control it gives over the experimental conditions. Well-defined constraints are imposed on the operators who key in the subjects' spoken responses, and all other aspects of the simulated system — the introduction of speech recognition errors, the dialogue processing, and the speech and other acoustic output — are entirely under software control.

Each call begins and ends with short periods of interaction between the WOZ operator and the subject. In the initial interaction, the operator delivers a predefined preamble to introduce the subject to the task, and makes sure that the subject is ready before initiating the dialogue with the simulated automated system. When the dialogue between the system and the subject ends, the operator is connected to the subject again for a brief signing-off conversation. The operator's microphone is disabled automatically while the subject interacts with the system to eliminate any possibility of interference with the dialogue.

The DTMF version of the system differs from the versions with speech input in that the user's inputs (keypresses) are handled automatically by the telephone interface card, which passes them on to the simulation software, rather than requiring listening and typing by the operator. In this case the role of the operator during the service dialogue is limited to monitoring the interaction and noting any anomalies that occur.

2.2. *Home shopping service*

The service simulated in this experiment was a home shopping service in which the user can order items from a catalogue by entering appropriate digit strings. The dialogue begins with a message welcoming the user and proceeds through three phases of data entry, in which the user is prompted to enter an eight-digit customer identification number, an eight-digit item number and a 16-digit credit card number. After entry of the customer number or credit card number, the number recognized by the system is read back and the user is asked for a yes/no response as to its correctness. After entry of the item number, the description of the item from the catalogue is read back, with its price, and again a yes/no confirmation response is requested from the user. If the number does not correspond to any item in the catalogue, a message stating this fact is played instead. In case of a negative response to the confirmation question, or an invalid item number, the current data entry phase is repeated. When the transaction is complete, the description and price of the item ordered are read back again, for information only.

In the IW dialogue, the user is instructed to enter each number one digit at a time, following the tones. Each number is presented in blocks of four digits; there is a spoken prompt at the beginning of each block, while the prompts between digits within a block consist only of tones. If the user speaks over the tone instead of after it, or says something outside the permitted vocabulary, which for digit entry consists of the words 'one' to 'nine', 'zero', 'nought' and 'oh', the operator hits a 'reject' key, and the system gives an error message ('Sorry, I didn't understand that') and prompts the user for the current digit again. In the case of no input from the user (and therefore no keystroke from the WOZ operator), the input times out after a predetermined period and the dialogue responds with a suitable prompt. After each block of four digits has been entered, the system reads back the recognized digits and asks whether they are correct; if the user says 'no', the system requests that block of digits again. This block-by-block confirmation is in addition to the confirmation of the whole number or item details at the end of each phase of the transaction.

The CW dialogue is similar except that each block of four digits is to be spoken connectedly after a single tone. If fewer than four digits have been entered by the end of the timeout period, or if any non-digit input is given, the input is rejected and the user is prompted for the block of digits again.

In the DTMF dialogue, there is a single prompt for each of the three numbers, which is to be entered as a single sequence of eight or 16 digits using the telephone keypad. In this version of the service no tone is played, and the user can 'type through', thereby interrupting the prompt. If an invalid input is given (for instance if the user presses the '*' key instead of a digit key), or if not enough digits have been entered by the end of the timeout interval, the input is rejected and the user is prompted for the number again. Confirmation of input is performed only at the end of each phase of the dialogue. The confirmation prompt asks the user to press '1' if the information read back is correct or '2' if it is incorrect.

In each case, up to three attempts to enter a digit or digit sequence are allowed. If the third attempt is rejected, the dialogue fails; the user is informed that there seems to be a problem, and is returned to the operator. Also, up to three attempts at each phase of the task are allowed: the dialogue fails if the user answers 'no' to the confirmation question (or enters an invalid item number) on three successive occasions.

The dialogue specifications were designed to be realistic with respect to current DTMF and speech recognition technology. In particular, the requirement to speak at a specified time (after the tone) in the IW and CW dialogues is a means of reducing the recognition error rate, since it is more difficult to detect and recognize spoken input if it occurs before the end of the system's speech output — whereas in the DTMF dialogue the user is allowed to interrupt the prompts because this is easier to implement reliably for DTMF and is a common feature of currently available automated services. Also the division of the numbers into four-digit blocks in the speech input cases is designed to reduce the incidence of dialogue failure due to errors and uncertainties in the speech recognition: there is a better chance of overcoming any recognition error when the recognition is checked and (if necessary) corrected by the user after every four digits than if correction depends on a successful entry of eight or 16 digits at once.

2.3. *Subjects*

Subjects were recruited from the Edinburgh area through a market research company. A set of 84 subjects participated in this experiment. The distributions of sex, age group and socio-economic grouping (HMSO 1991) are shown in table 1.

2.4. *Procedure and experimental task*

Subjects attended the research centre at the University of Edinburgh individually. On arrival, they first completed the NEO PI-R five-factor test (Costa and McCrae 1992) consisting of 240 personality questions.

Table 1. Analysis of subjects by sex, age and socio-economic grouping. HMSO Scottish categories I to V correspond to categories A to E in England and Wales, representing professional, intermediate, skilled, partly skilled and unskilled occupations respectively.

Sex		Age group		Socio-economic grouping	
Gender	Number	Group	Number	I – II	III – V
Male	38	18 – 38	16	4	12
		39 – 59	10	6	4
		60+	12	2	10
Female	46	18 – 38	16	4	12
		39 – 59	20	3	17
		60+	10	4	6

This test, which is not timed but typically takes between 45 minutes and an hour to complete, allowed scores to be calculated for each subject on five personality traits: neuroticism, extraversion, openness, agreeableness and conscientiousness.

Immediately following this, subjects completed the AH4 Group Test of General Intelligence (Heim 1970) involving two 10-minute timed tests, one for spatial ability and one for verbal ability. These are referred to here as cognitive skills.

After a short break, subjects used the home shopping service three times, each time with a different input mode: connected word (CW), isolated word (IW) and DTMF. Order of exposure to the versions was randomized in such a way that 14 subjects experienced each of the six possible orderings. The procedure for each use of the service was as follows.

1. The experimenter gave the subject a brief description of the home shopping service and the task to be accomplished with it, including a demonstration of the data entry method to be used. The subject was provided with the following printed reference materials:

   - a sheet of paper giving the customer identification number;
   - a card bearing the credit card number to be used in the dialogue;
   - a copy of the catalogue, consisting of four pages (one sheet of A3 paper folded to A4 format), and containing descriptions, pictures and reference numbers for the 14 items available for ordering.

   The customer number and credit card number were the same for all uses of the service by any one subject, but the item to be ordered, which was specified in the task description given by the experimenter, was different each time.
2. The experimenter went to another room ‘to start up the system’; this in fact entailed simply telling the WOZ operator to initiate the call. When initiated by the operator, the software called the number of the telephone that the subject was to use, and the operator proceeded with the preamble when the subject answered. Once the call had begun, the experimenter returned to the subject.
3. The subject used the service while seated at a table (so that the printed reference materials could be turned over easily with one hand).

Following each use of the service, subjects completed a Likert attitude and usability questionnaire. As a result, mean attitude scores on a seven-point scale were available for each subject and for all three data entry modes. Comparison of the results gives a subjective rank ordering of input modes by each subject.

After these three uses of the home shopping service, subjects were asked to select their most preferred version for a fourth use. The procedure for the fourth use was similar to that for the first three uses. This part of the experiment provided a behavioural measure of the subjects’ preferences.

Finally, subjects were interviewed for about 10 minutes regarding the different versions of the service they had used as well as their familiarity with and attitudes towards automated telephone services, computers and technology in general. One of the interview questions asked them to identify their least preferred version of the service they had just used. This piece of information, taken together with their earlier choice of most preferred version, allowed the construction of an objective rank order of preference for each subject which is to be contrasted with the subjective rank ordering arising from the attitude questionnaire scores.

2.5. *Questionnaire*

The questionnaire completed after each use of the service asked for responses on a seven-point Likert scale (Oppenheim 1966), from ’strongly agree’ to ’strongly disagree’, to 32 statements about the version of the service the user had just experienced. The statements were designed to cover the set of generic attributes of automated telephone services identified in Dutton *et al.* (1993) as well as the key attributes of the input mode (speech or keypad, isolated or connected). They were balanced for polarity so that there were equal numbers of statements indicating positive and negative evaluations of the service. Most of the statements were the same for all three versions (DTMF, IW and CW), but two statements (about the tone) were not applicable to the DTMF version and were therefore omitted in this case, and three others varied in their wordings to suit the different input modes.

The questionnaire responses were converted into numerical values on a seven-point attitude scale, from 1 (most unfavourable) to 7 (most favourable), allowing for the polarities of the statements. Thus, for instance, a ‘strongly agree’ response to a negative statement would be converted to a value of 1. Each subject’s overall attitude to a version of the service could then be measured by taking the mean of these numbers across all the items in the questionnaire.

3. Results

3.1. *Rank ordering — expressed user preferences*

The number of subjects selecting each of the six possible rank orderings of the three input modes is shown in table 2.

The preferences are heavily biased in favour of both DTMF and CW. Only one subject chose IW as the most preferred option while 22 and 61 subjects designated IW as their second and third choice respectively.

Overall, 45 subjects (53.6%) most preferred DTMF (rows 1 and 2), while 38 (45.2%) most preferred CW (rows 5 and 6), indicating a slight overall preference in objective rank ordering for DTMF.

Within the groups most preferring CW and DTMF there is a difference in distribution with respect to the least preferred option. In the case of those rating DTMF as most preferred, nearly all (91%) rated IW as their least preferred mode. In the case of those rating CW as most preferred, there was an approximately equal split between IW and DTMF as the least preferred mode. Thus there is a group of subjects who show low preference for DTMF and rate it even below IW. This group of subjects is shown on row 6 of table 2 and amounts to 21.4% of the experiment sample.

Table 2. Objective rank ordering of data entry modes.

Most preferred		Least preferred	Number of subjects
DTMF	IW	CW	4
DTMF	CW	IW	41
IW	DTMF	CW	1
IW	CW	DTMF	0
CW	DTMF	IW	20
CW	IW	DTMF	18

3.2. *Attitudes and subjective rank ordering*

Since each subject completed three attitude questionnaires, one for each of the data entry modes, it is possible to obtain a rank ordering based on the questionnaire means. This is referred to here as the subjective rank ordering. The results are shown in table 3 and can be compared with the objective ordering results shown in table 2.

As in table 2, each subject is here taken to contribute 1 to the total count. However, some subjects returned the same mean for two of their questionnaires. In order to maintain the same total as in table 2 and to avoid losing information, each of these subjects is taken as making contributions of 0.5 to two rows in the table, hence the fractional counts of subject numbers.

As in the case of the objective ranking, the distribution of preferences is biased in favour of DTMF and CW against IW. A total of 10.5 subjects ranked IW as their most preferred option while 21.5 ranked it second and 52 ranked it as least preferred. A total of 45.5 subjects (54.2%) most preferred DTMF and 28 (33.3%) most preferred CW. CW was therefore ranked less highly subjectively than it was objectively; IW, in contrast, was ranked more highly subjectively than objectively.

Table 3. Subjective rank ordering of data entry modes.

Most preferred		Least preferred	Number of subjects
DTMF	IW	CW	10
DTMF	CW	IW	35.5
IW	DTMF	CW	3.5
IW	CW	DTMF	7
CW	DTMF	IW	16.5
CW	IW	DTMF	11.5

The overall mean attitude score for the DTMF service was 5.60, for the CW version 5.47 and for the IW version 5.09. Pairwise related-sample t-tests showed highly significant differences between DTMF and IW and between CW and IW ($p < 0.001$ in each case), but no significant difference between DTMF and CW.

These results are based on the 30 questionnaire items which are directly comparable across the three versions of the service: the responses to the two statements about the tone prompt, found only in the IW and CW questionnaires, have been omitted from the calculations. With all 32 questionnaire items included, the mean attitude score for CW was 5.44, and for IW 5.04. The drop in the scores with the full set of statements is due mainly to the relatively negative attitudes expressed in responses to the statement ‘I didn’t like the beep’.

3.2.1. *Subjective rank ordering as a predictor of objective rank ordering:* Given that this experiment produced data on both subjective ranking (questionnaire) and objective ranking (behaviour), it is possible to measure the relationship between the two rankings, and in particular to evaluate how effective the questionnaire is in predicting user preferences as expressed in objective behaviour. This is a measure of the objective validity of the attitude questionnaire.

For 58 of the 77 subjects whose questionnaires yielded a single best-ranking input mode, this mode was the one chosen explicitly as most preferred. Of the other seven subjects, five chose one of the two modes rated first equal in their questionnaires. Thus, counting the prediction success as 1 for the former subjects and 0.5 for the latter ones, the prediction of behaviour (choice of most preferred mode) from questionnaire responses was 72% accurate.

Analysis of the results reveals a marked variation in prediction accuracy across input modes. None of the 10.5 subjects who scored IW most highly in their questionnaires chose this mode for the fourth use of the service, whereas 23.5 (84%) of the 28 scoring CW most highly and 37 (81%) of the 45.5 scoring DTMF most highly selected the same mode for the fourth use.

The accuracy of prediction of the explicit choice of least preferred mode from the subjective rankings was 70%; the rate of correct prediction of the full objective rank order was 55%.

The correspondence between the questionnaire results and the expressed preference ranking is certainly better than random, and so provides a degree of validation of the questionnaire. However, the mismatch between the subjective and objective rankings of the IW service for some subjects suggests that further investigation and perhaps refinement of the subjective attitude measurement (for instance to use a weighted mean of the responses to the questionnaire statements) might be desirable.

3.3. *Order of use and user attitudes*

It may be the case that attitude to one input mode is affected by exposure to other modes. Since in this experiment each subject experienced all three modes, the presence of such an interaction can be investigated. Figure 2 shows attitude questionnaire means against order of exposure for each input mode.

One-way ANOVAs taking order of use as the categorical variable and mean attitude as the dependent variable for each input mode showed no significant effect of order of use on attitude to CW or DTMF. There was however a significant effect of order of use on attitude to IW ($p < 0.01$).

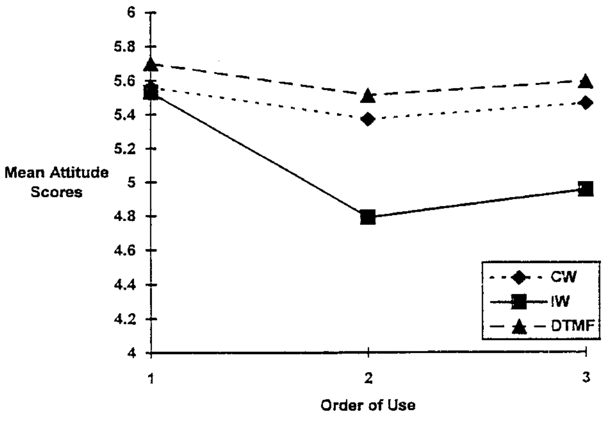

Figure 2. Mean attitude score (based on 30 questionnaire items) against order of use.

This result is particularly interesting both in corroborating the unexpected result obtained from the previous experiment (mentioned in section 1) and in its implications for the design of services for novice and experienced users. IW received the same mean attitude score for novice users as CW and only slightly lower than DTMF. This indicates that users with no previous experience of automated telephone services adopt very similar attitudes irrespective of input mode, a fact which explains the lack of difference in attitude found towards CW and IW modes in the previous experiment. However, as soon as subjects had experience of different input modes, IW was judged to be significantly less usable than CW or DTMF.

Further analysis showed no significant difference in the attitudes of the subjects using IW as their second version of the service between those who had used the CW version first and those who had used the DTMF version.

These results offer a partial explanation for the observation in section 3.2 that IW was more highly ranked relative to the other modes in the questionnaire responses than in the objective choice. Subjects completed the questionnaire for each version of the service immediately after using it, and therefore in some cases without having experienced any other version, whereas the explicit choice amongst the versions was made after experience of all three. However, this does not account fully for the difference in the subjective and objective rankings, since only five of the 10.5 subjects for whom IW was first in the subjective ranking (but not chosen objectively as most preferred) experienced the IW mode first.

3.4. *Cognitive skills*

The AH4 Group Test of General Intelligence is standardly given in two parts: the first part tests verbal ability, and the second tests spatial ability. The positive correlation between the two sets of scores across the subjects in this experiment was high (Pearson $r = 0.78$), a finding typical of the population at large.

For the purpose of carrying out ANOVAs based on verbal or spatial ability as the categorical variable, the scores were converted to a binary distinction between *high* and *low* for each test, the former being defined as the upper half of the sample distribution, the latter being the lower half of the sample distribution.

3.4.1. *Verbal abilities:* Figure 3 shows the mean attitudes to CW, DTMF and IW input modes for low and high verbal abilities.

The attitudes for CW and IW were negatively correlated with verbal ability, while the attitude towards DTMF was positively correlated with verbal ability. However, ANOVAs showed that the difference in attitude between the low and high ability groups was not significant for any of the input modes.

3.4.2. *Spatial abilities:* Figure 4 shows the mean attitudes to CW, DTMF and IW input modes for low and high spatial abilities.

This showed a similar pattern to verbal ability, but with two important differences. Firstly, the mean attitude to DTMF was negatively correlated with high spatial ability. Secondly, the ANOVA result showed a highly significant difference between the low and high spatial ability groups for CW ($p < 0.005$).

3.4.3. *Cognitive skills and relative attitudes:* The analyses above explored the effects of verbal and spatial ability on attitudes to each input mode taken singly. Further analysis was carried out on the subjects' *relative* attitudes to the different input modes, as measured by the per-subject between-mode attitude differences, to investigate how the preferences between modes of input varied with the level of cognitive skill.

Within each group of subjects (high or low ability, verbal or spatial), related-sample *t*-tests showed significant attitude differences ($p < 0.01$) between DTMF and IW and between CW and IW, with attitudes to the IW service being consistently lower. For subjects with high cognitive skills (verbal or spatial), there was also a significant difference ($p < 0.01$) between DTMF and CW. ANOVAs on the relative attitudes showed that the difference in the DTMF-CW relative attitude between the low and high ability groups was significant ($p < 0.005$), whether based on verbal or spatial ability; the differences in the other relative attitudes (DTMF-IW and CW-IW) across ability levels did not reach significance. The pattern can be seen in figures 3 and 4 where the divergence between the means for DTMF and CW is evident in the high ability groups.

These results may be summarized as follows. Firstly, all subjects, whether of low or high cognitive skills, rated IW significantly worse than either DTMF or CW. Secondly, subjects high in cognitive skills significantly preferred DTMF over CW. Thirdly, the *difference* in the DTMF-CW preferences between the low and high ability subjects was significant.

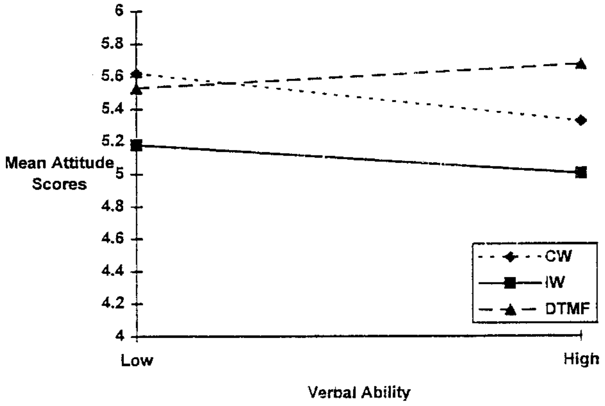

Figure 3. Mean attitude score (based on 30 questionnaire items) against subject's verbal ability level.

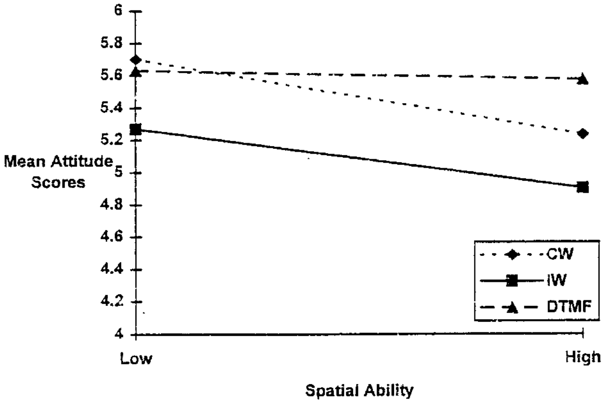

Figure 4. Mean attitude score (based on 30 questionniare items) against subject's spatial ability level.

3.5. *Personality*

For each of the five personality factors measured by the NEO PI-R test, the subjects were divided into high and low scoring groups as in the case of the cognitive skills, and ANOVAs were performed with the per-mode attitudes and pairwise relative attitudes as dependent variables.

The only significant result was for the factor *agreeableness*. The low-agreeableness subjects had a significantly higher preference for DTMF relative to CW than the high-agreeableness subjects ($p < 0.05$).

3.6. *Other subject characteristics*

For each input mode, a one-way ANOVA was performed with age group (as shown in table 1) as the categorical variable and attitude as the dependent variable. No significant variation of attitude with age was found for any of the three input modes.

A similar series of ANOVAs was carried out with the sex of the subject as the categorical variable. Again no significant difference was found for any of the input modes. The mean attitudes of male and female subjects were very similar for each input mode.

3.7. *Performance measurements*

Although the main focus of this experiment was on users' preferences, two aspects of the subjects' performance in using the service were also measured and were examined for effects of input mode and user characteristics — namely *call duration* and *number of user errors*.

3.7.1. *Call duration:* The mean call duration for all subjects when using DTMF was 114 s; when using CW, 156 s; and when using IW, 173 s. The longer times taken with the speech input modes reflect not only any differences in the inherent speed of data entry but also the differences in the dialogues, which involved more prompts and confirmations for IW and CW than for DTMF.

ANOVAs revealed significant differences between the low and high verbal ability groups for CW ($p < 0.01$) and for DTMF ($p < 0.001$), but not for IW. The results for the low and high spatial ability groups were similar though the significance for CW was lower ($p < 0.05$). In each case the mean DTMF or CW call duration was shorter for the high-ability group.

A significant difference among the age groups was found for DTMF ($p < 0.001$), but not for CW or IW. The mean DTMF call duration was greater for the 60+ group (127 s) than for the younger groups (108 s and 110s).

No significant call duration differences were found for sex or personality.

3.7.2. *User errors:* The mean number of user errors (missing or invalid inputs) per call was 0.27 for DTMF, 0.33 for CW and 0.62 for IW. Related-sample t-tests showed significant differences between the error counts for DTMF and IW ($p < 0.01$) and between those for CW and IW ($p < 0.05$).

For each input mode, one-way ANOVAs were performed for the numbers of user errors with the various individual characteristics as categorical variables. No significant effects were found for verbal or spatial ability, personality, age or sex for any of the input modes.

4. Discussion

4.1. *General remarks*

Conclusions may be drawn from a study of this kind at different levels of generality.

Results as to the *overall preferences amongst input modes* have a direct practical application in determining the choice of data entry mode, provided that the service dialogue in view is sufficiently similar to the one used in the experiment. However, caution is in order in extrapolating from these results since, as mentioned in section 1, the choice of input mode may vary with the type of data to be entered, with the dialogue strategy and with the level of recognition accuracy available for speech input.

Specifically, this was a best-case comparison in the sense that 100% accuracy was adopted for both speech input modes; the results for IW and CW are therefore upper bounds on what would be attained with real recognizers and the same dialogue design. The choice of 100% accuracy here was dictated by the background mentioned in section 1: the previous experiment had shown no advantage of CW over IW mode at 100% and it was desired to investigate this unexpected finding before proceeding to lower accuracies. Further experiments with both speech input modes at lower accuracies would be desirable, now that a difference has been established at 100% to see at what point the preference switches in favour of IW because of the more severe effect of imperfect recognition technology in the CW case. (Jack *et al.* (1992) reported results at a range of recognition accuracies, for IW only, with a simple digit entry task. At 100% accuracy the mean attitude score was 5.88; this fell to 5.27 at 95%, 4.99 at 90% and 4.96 at 85%. The percentages here are nominal accuracies of effective digit recognition, after querying or reprompting in cases of uncertain initial recognition; the true effective accuracies were about 3% lower, except in the 100% case, because of occasional errors in recognizing 'yes' and 'no'.)

While the use of 100% accuracy throughout would tend to make this comparison unrealistically favourable towards speech, there were also features of the dialogue designs which favoured DTMF. The DTMF dialogue allowed interruption of prompts and required less confirmation of entered data than the speech input dialogues. With accurate enough speech recognition and echo cancelling technology, it would be possible to introduce these advantages into the speech input versions of the service as well.

There are also practical issues in the use of DTMF which fall outside the scope of the present experiment, arising from the fact that some users do not have DTMF telephones, and from the adverse affect in other cases of having the keypad on the handset which makes the combination of listening and keying difficult.

In contrast, the *differences* in preference observed between novice and experienced users and between subjects with different personal characteristics can be expected to hold more generally, in that the directions of the effects due to these factors will be the same across a wide range of applications and dialogues.

4.2. *Preferences between DTMF and speech*

An important finding in this experiment is the existence of a substantial minority of subjects who did not like DTMF as an input mode. The tasks in this simulated home shopping service required the entry of digits only (except for yes/no confirmations). It would seem therefore that they were well suited to DTMF input. Nevertheless, 21% of the subject sample identified DTMF as their least preferred mode, ranking it even below IW, which on the whole was rated poorly in this experiment. This suggests that in the development of new automated services, people who might choose not to use DTMF ought to be taken into account by offering speech input (preferably CW if accurate enough) as an alternative either directly or via multimodal options.

4.3. *Order effects*

The order effects explored in this experiment are of particular interest both for their consequences for future experimental designs and for their implications with regard to the development of automated services.

The experiment established that all three input modes are similarly well rated by novice users, a fact which explains the results obtained in the previous experiment where the attitudes to CW and IW at 100% recognition accuracy were almost identical and no statistically significant difference could be established between them.

When users have experience of more than one input mode, however, they rate IW substantially worse than either DTMF or CW. This result has important consequences for the implementation of services, as it suggests that users who are already familiar with either a CW or a DTMF service will adopt a more negative attitude towards a new service offering IW input only.

4.4. *Cognitive skills*

Verbal and spatial reasoning skills are a variable in the potential user population of any automated telephone service. It is therefore of interest to know whether differences in these skills interact with attitudes to services for different input modes. This issue can be addressed in different ways depending upon the perspective adopted.

For example, it might be proposed to target an automated telephone service at a user population which is believed to be very diverse with regard to the levels of cognitive skills of its members. If the service designers are constrained to adopt a specific data input mode (for example, CW because of the lack of availability of DTMF telephones), they will want to know if there is a relationship between level of cognitive skill and attitude towards the service for that input mode. Results reported here suggest that attitude to CW input mode is affected by level of cognitive skill whereas attitudes to IW and DTMF are less affected.

On the other hand, a service provider may have in mind a specific user population which is fairly homogeneous with regard to levels of cognitive skill. In this case, the service designers will want to know the most appropriate input technology to use in order to provide the best level of satisfaction for that population. Results indicate that for both groups, attitudes to IW are lower than attitudes to DTMF or CW and that, in addition, the group with high cognitive skills shows a better attitude to DTMF than to CW. While the overall balance of preference between DTMF and speech input may vary with the type of transaction and the form of the data to be entered, the tendency for high-ability users to be more favourable to DTMF relative to CW than those of lower ability seems likely to persist.

Acknowledgements

The authors wish to acknowledge the financial support for this research from BT under its Strategic University Research Initiative, and the helpful contributions to the planning of the experiments made by members of the research staff at BT Laboratories, Martlesham.

References

Bossemeyer, R.W. Jr. and Schwab, E.C. 1991, Automated alternate billing services at Ameritech: speech recognition performance and the human interface, *Speech Technology*, **5,** 3, 24 – 30.

Brooks, C.M. 1989, User interface design for speech recognition telephone applications, *Speech Technology*, **4,** 4, 58 – 60.

Costa, P.T. and McCrae, R.R. 1992, *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI) Professional Manual,* (Florida: Psychological Assessment Resources, Inc).

Dutton, R.T., Foster, J.C., Jack, M.A. and Stentiford, F.W. 1993, Identifying usability attributes of automated telephone services, *Proceedings of Eurospeech 93*, Berlin, 1335 – 1338.

Fay, D. 1993, Interfaces to automated telephone services: do users prefer TouchTone or automatic speech recognition, *Proceedings of the 14th International Symposium on Human Factors in Telecommunications*, Darmstadt, 339 – 348.

Foster, J.C., Dutton, R.T., Jack, M.A., Love, S., Nairn, I.A., Vergeynst, N.A. and Stentiford, F.W.M. 1992, Design and evaluation of dialogues for automated telephone services, *Proceedings of the Institute of Acoustics*, **14,** 6, 629 – 635.

Foster, J.C., Dutton, R.T., Jack. M.A., Love, S., Nairn, I.A., Vergeynst, N.A. and Stentiford, F.W.M. 1993, Intelligent dialogues in automated telephone services, in C. Baber and J.M. Noyes (eds), *Interactive Speech Technology: Human Factors Issues in the Application of Speech Input/Output to Computers* (London: Taylor and Francis), 167 – 175.

Fraser, N.M. and Gilbert, G.N. 1991, Simulating speech systems, *Computer Speech and Language*, **5,** 2, 81 – 99.

Hammond, G.T. 1990, Deploying speech technology in telephone banking services, *Speech Technology*, **5,** 2, 30 – 33.

Heim, A.W. 1970, *AH4 Group Test of General Intelligence Manual* (Windsor: NFER-Nelson Publishing).

HMSO 1991, *Office of Population and Census Surveys* (London: HMSO).

Hornstein, T. 1994, Telephone voice interfaces on the cheap, *Proceedings of the UBILAB'94 Conference*, Zurich, 134 – 147.

Jack, M.A., Foster, J.C. and Stentiford, F.W. 1992, Intelligent dialogues in automated telephone services, *Proceedings of ICSLP 92 (International Conference on Spoken Language Processing)*, Banff, Alberta, Canada, 715 – 718.

Karis, D. and Dobroth, K.M. 1991, Automating services with speech recognition over the public switched telephone network: human factors considerations, *IEEE Journal on Selected Areas in Communications*, **9,** 574 – 585.

Lea, W.A. 1980, The value of speech recognition systems, in W.A. Lea (ed.), *Trends in Speech Recognition* (Englewood Cliffs, NJ: Prentice-Hall), 3 – 18.

Levas, S. 1988, Automation potential of Operator Number Identification (ONI) intercept services, *Speech Technology*, **4,** 2, 26 – 29.

Martin, M.M., Williges, B.H. and Williges, R.C. 1990, Improving the design of telephone-based information systems, *Proceedings of the Human Factors Society 34th Annual Meeting,* 198 – 202.

Oppenheim, A.N. 1966, *Questionnaire Design and Attitude Measurement* (London: Heinemann).

Roberts, T.L. and Engelbeck, G. 1989, The effects of device technology on the usability of advanced telephone functions, *Proceedings of CHI'89*, 331 – 337.

Rosenbeck, P. and Baungaard, B. 1992, Experiences from a real-world telephone application: teleDialogue, *Proceedings of ICSLP 92 (International Conference on Spoken Language Processing)*, Banff, Alberta, Canada, 1585 – 1588.

Schumacher, R.M. Jr. 1992, Phone-based interfaces: research and guidelines, *Proceedings of the Human Factors Society 36th Annual Meeting*, 1051 – 1055.

Waterworth, J. 1984, Speech communication: how to use it, in A. Monk (ed), *Fundamentals of Human-computer Interaction* (London: Academic Press), 221 – 236.

Wilpon, J.G., Mikkilineni, R.P., Roe, D.B. and Gokcen, S. 1990, Speech recognition: from the laboratory to the real world, *AT&T Technical Journal*, **69,** 14 – 24.