

The design of speech-based automated mobile phone services using interface metaphors

A thesis submitted for the degree of Doctor of Philosophy

By

Mark David Howell

**Department of Information Systems and Computing,
Brunel University**

December 2004

ABSTRACT

Interface metaphor is a widely used design technique for interactive computer systems. The advantages of using interface metaphors derive from their ability to promote active learning, which enables a user to transfer knowledge from a familiar real world domain, to an unfamiliar computing domain. Interface metaphor is not currently used for the design of automated phone services, and it was the aim of this thesis to examine whether interface metaphor could improve the usability of speech-activated automated mobile phone services. A human-centred design methodology was followed to generate, select, and develop potential metaphors, which were used to implement metaphor-based phone services. An experimental methodology was then used to compare the usability of the metaphor-based services with the usability of currently available number-based phone services. The first experiment examined the effect of three different interface metaphors on the usability of a mobile city guide service. Usability was measured as a range of performance and attitude measures, and was supplemented by telephone interview data. After three consecutive days of usage, participants both preferred, and performed better with, the service that was based on an office filing system metaphor. Experiment two was conducted over a six week period, and investigated the effect of users' individual differences, and the context of use, on the usability of both the office filing system metaphor-based service, and a non-metaphor service. The results showed that performance with the metaphor-based service was significantly better than performance with the non-metaphor service. The usability of the metaphor-based service was not significantly affected by users' individual characteristics and aptitudes, whereas the number-based service was, suggesting that metaphor-based services may be more usable for a wider range of potential users. Usability levels for both services were found to be consistent across both private and public locations of use, suggesting that speech-activated mobile phone services provide a flexible means of information access. Experiment three investigated the strategies used by participants when interacting with mobile phone services, specifically the visualisation strategy that was used by two thirds of the metaphor-based service participants in experiment two. In addition to the attitude and performance measures used for experiments one and two, face-to face interviews were conducted with participants. The results indicated that significantly more participants visualised the metaphor-based services relative to a non-metaphor service, and that visualisation of the service structure led to significant performance improvements. This thesis has demonstrated the usability benefits of interface metaphor as a design technique for speech-based mobile phone services. These benefits of metaphor appear to derive from their ability to provide a mental model of the phone service that can be visualised, and their ability to accommodate the individual differences of users.

TABLE OF CONTENTS

Abstract	i
Table of Contents	ii
List of Figures	x
List of Tables	xi
Acknowledgements	xvii
Declaration	xviii
Chapter 1. Introduction	1
1.1 Introduction.....	1
1.2 Research motivation.....	2
1.3 Research methods.....	5
1.4 Thesis overview	6
Chapter 2. Literature review	8
2.1 Introduction.....	8
2.2 Automated phone services	9
2.2.1 Introduction to automated phone services.....	9
2.2.1.1 Mode of communication	11
2.2.1.2 Dialogue design.....	13
2.2.2 Conversational automated phone services	17
2.2.2.1 Components of human conversation	19
2.2.2.2 Limitations of speech-recognition technology.....	21
2.2.3 Usability problems of automated phone services	23
2.2.3.1 Additional dialogue.....	24
2.2.3.2 Earcons.....	25
2.2.3.3 Conversational interfaces	26
2.2.4 Implications for this research.....	28
2.3 The role of metaphor in user interface design.....	29
2.3.1 Introduction to metaphor.....	29
2.3.2 Interface metaphors.....	33
2.3.3 Mental models.....	35

2.3.4	The metaphor design process	39
2.3.5	Empirical investigations of metaphor in computing	46
2.3.6	Categories of metaphor	49
2.3.7	Interface metaphors for different computing paradigms.....	52
2.3.8	Problems of metaphor	56
2.3.9	Alternatives to metaphor.....	61
2.3.10	Implications for this research	63
2.4	Individual differences in Human-Computer Interaction	64
2.4.1	Introduction.....	64
2.4.1.1	Age	67
2.4.1.2	Gender	69
2.4.1.3	Prior telephone and computing experience	70
2.4.1.4	Verbal ability.....	71
2.4.1.5	Spatial ability	73
2.4.1.6	Cognitive style	76
2.4.1.7	Working memory	78
2.4.1.8	Attitude towards mobile phone usage in public	79
2.4.2	Implications for this research	81
2.5	Context of use	82
2.6	Conclusions from the literature review	84
Chapter 3.	Research Methodology	86
3.1	Introduction.....	86
3.2	Overview of the programme of research.....	86
3.2.1	Methodology One: Human-centred design	87
3.2.1.1	Stage One: Planning the human-centred design process....	89
3.2.1.2	Stage Two: Understand and specify the context of use	90
3.2.1.3	Stage Three: Specify user and organisational requirements.....	90
3.2.1.4	Stage Four: Produce design solutions	91
3.2.1.4.1	Wizard of Oz prototyping	93
3.2.1.5	Stage Five: Evaluation	98
3.2.2	Methodology Two: Experimental methodology	100
3.2.2.1	The experimental plan.....	102

3.2.2.2 Variables	102
3.2.2.2.1 Operational definitions for the subjective variables	103
3.2.2.2.2 Operational definitions for the objective variables	105
3.2.2.3 Design	108
3.2.2.4 Sample of participants.....	109
3.2.2.5 Data collection instruments.....	111
3.2.2.5.1 Usability questionnaire	111
3.2.2.5.2 TrueActive monitoring software.....	113
3.2.2.5.3 Interviews.....	113
3.2.2.6 Data analysis	114
3.3 Chapter summary	115

Chapter 4. Designing interface metaphors for automated mobile phone

services	116
4.1 Introduction	116
4.2 Preliminary Study 1: Generating, selecting, and developing interface metaphors	117
4.2.1 Introduction.....	117
4.2.1.1 Generating interface metaphors – brainstorming	117
4.2.1.2 Selecting interface metaphors – card sorting and sketching	118
4.2.2 Methodology	120
4.2.3 Results.....	122
4.2.3.1 Usability of the methodologies	122
4.2.3.2 Productivity of the methodologies	122
4.2.3.3 Using metaphors to explain an automated telephone service	123
4.2.3.4 Card sorts, sketches, and their explanations	124
4.2.3.5 POPITS features used in the explanations.....	131
4.2.3.6 Metaphor categories.....	131
4.2.4 Discussion	133

4.3 Preliminary Study 2: Improving the usability of mobile phone services using spatial interface metaphors	134
4.3.1 Introduction	134
4.3.2 Prototype design and development	135
4.3.3 Methodology	142
4.3.3.1 Design	142
4.3.3.2 Participants.....	142
4.3.3.3 Apparatus	142
4.3.3.4 Data collection	143
4.3.3.5 Procedure.....	144
4.3.4 Results	144
4.3.4.1 Subjective measures	144
4.3.4.2 Objective measures	145
4.3.5 Discussion	146
4.4 Chapter summary	147
Chapter 5. Spatial metaphors for a speech-based mobile city guide service.....	149
5.1 Introduction	149
5.2 Prototype design and development	150
5.3 Methodology	154
5.3.1 Design	154
5.3.2 Participants.....	154
5.3.3 Apparatus	155
5.3.4 Data collection	155
5.3.4.1 Subjective measures	155
5.3.4.2 Objective measures	156
5.3.5 Procedure	156
5.4 Results.....	158
5.4.1 Subjective attitude measures	159
5.4.1.1 Differences between groups.....	161
5.4.1.2 Differences between trials.....	163
5.4.2 Objective performance measures	163
5.4.2.1 Differences between groups.....	165
5.4.2.2 Differences between trials	167

5.5 Discussion	167
5.6 Chapter summary	170

Chapter 6. The impact of interface metaphor and context of use on the usability of a speech-based mobile city guide service173

6.1 Introduction	173
6.2 Prototype design and development	175
6.3 Methodology	176
6.3.1 Design	176
6.3.2 Participants.....	178
6.3.3 Apparatus	178
6.3.4 Data collection	178
6.3.5 Procedure	182
6.4 Results.....	183
6.4.1 The effect of metaphor on performance and attitude.....	184
6.4.2 The effect of context on performance and attitude	187
6.4.3 Post task interviews.....	189
6.4.4 Multiple regression models	190
6.4.4.1 Performance measures	190
6.4.4.2 Attitude measures.....	194
6.4.4.3 Multiple regression summary.....	197
6.5 Discussion	197
6.6 Chapter summary	201

Chapter 7. The effect of visualisation on the usability of voice-operated metaphor-based mobile phone services.....203

7.1 Introduction	203
7.2 Prototype design and development	205
7.3 Methodology	206
7.3.1 Design	206
7.3.2 Participants.....	206
7.3.3 Apparatus	207
7.3.4 Data collection	207
7.3.5 Procedure	208

7.4 Results	209
7.4.1 Participant variables	209
7.4.2 Performance measures	210
7.4.3 Attitude measures	211
7.4.4 Recall of service structure	212
7.4.5 Visualisation of the service structure	213
7.4.5.1 Performance measures	213
7.4.5.2 Attitude measures	214
7.4.6 Qualitative data	215
7.4.7 Design factors	224
7.5 Discussion	227
7.6 Chapter summary	233
Chapter 8. Conclusions	235
8.1 Summary of the experimental findings	235
8.1.1 Computer component: Interface metaphor	236
8.1.2 Human component: Individual differences	241
8.1.3 Environment component: Context of use	243
8.2 Limitations and future directions	244
8.3 Final reflections	247
References	248
Appendices	290
Appendix 1. Likert questionnaire statements matched to the 6 subjective factors ...	290
Appendix 2. The Likert style usability questionnaire used for experiments one, two, and three	292
Appendix 3. The telephone Internet service structure	294
Appendix 4. The telephone city guide service structure	295

Appendix 5. The 60 ‘real world systems’ generated in the brainstorming session...	296
Appendix 6. The POPITS sheet	298
Appendix 7. The Likert style usability questionnaire used for preliminary study one.....	300
Appendix 8. The metaphor list and ranking scores from preliminary study one	301
Appendix 9. The Likert style usability questionnaire used for preliminary study two.....	303
Appendix 10. The Technographic questionnaire	303
Appendix 11. The Principal Components Analysis study conducted to evaluate the construct validity of the Technographic questionnaire	308
Appendix 12. The task sheet for preliminary study two	318
Appendix 13. The task sheet for experiment one.....	319
Appendix 14. The post-task informal interview questions for experiment one.....	321
Appendix 15. The post-task informal interview questions for experiment two.....	322
Appendix 16. The mobile phone attitude questionnaire	323
Appendix 17. The task sheet for experiment two.....	325
Appendix 18. The post-task informal interview questions for experiment three.....	327
Appendix 19. The multiple choice memory questionnaire	328
Appendix 20. The task sheet for experiment three.....	330

Appendix 21. Absolute transaction times331

LIST OF FIGURES

Figure 1.1. The three-level model of HCI (Eason, 1991)	4
Figure 2.1. The design model, the user's model, and the system image (Norman and Draper, 1986, p. 46)	38
Figure 2.2. The interaction of the metaphor and the system in HCI	42
Figure 3.1. Human-centred design activities (from ISO 13407)	88
Figure 4.1. A card-sort of a road system metaphor	125
Figure 4.2. A card-sort of a filing cabinet metaphor	126
Figure 4.3. A card-sort of a human circulatory system metaphor	127
Figure 4.4. A sketch of a brain metaphor	128
Figure 4.5. A sketch of a supermarket metaphor	129
Figure 4.6. A sketch of a filing cabinet metaphor	130
Figure 4.7. One of the 3 main branches of the city guide service menu hierarchy ...	136
Figure 5.1. Mean scores across experimental trials for: Cognitive Demand	162
Figure 5.2. Mean scores across experimental trials for: Time as a Percentage of Total Prompt Time	166
Figure 6.1. Mean scores across experimental trials for time	185
Figure 6.2. Mean scores across experimental trials for prompt interrupt	186
Figure 6.3. Mean scores across experimental trials for nodes	186

LIST OF TABLES

Table 2.1. POPITS features for a Book metaphor.....	45
Table 3.1. Methods for human-centred design.....	89
Table 3.2. The design and evaluation methods used for the experimental work	93
Table 3.3. The 8 performance measures recorded for experiments one, two, and three.....	106
Table 4.1. Usability questionnaire results from preliminary study one	122
Table 4.2. POPITS features for the ‘Shopping Centre’ metaphor.....	122
Table 4.3. POPITS results for the card sorting and sketching methodologies.....	123
Table 4.4. Top 10 metaphors for each telephone service.....	131
Table 4.5. The 5 metaphor categories and their descriptions.....	132
Table 4.6. The 3 highest scoring metaphor categories, and metaphors, from preliminary study one.....	133
Table 4.7. Features of the standard service used for preliminary study two	139
Table 4.8. Features of the travel system service used for preliminary study two	139
Table 4.9. Features of the office filing system service used for preliminary study two.....	139
Table 4.10. Features of the shopping system service used for preliminary study two.....	140

Table 4.11. Example dialogue from the 4 services used for preliminary study two	141
Table 4.12. Descriptive data for the individual difference measures	144
Table 4.13. Paired t-test results for attitude between the metaphor and non-metaphor services	145
Table 4.14. Paired t-test results for ‘time’ between the metaphor and non-metaphor services	145
Table 5.1. Features of the standard service used for experiment one	153
Table 5.2. Features of the travel system service used for experiment one	153
Table 5.3. Features of the office filing system service used for experiment one	153
Table 5.4. Features of the shopping service used for experiment one	154
Table 5.5. Attitude measures for experiment one	155
Table 5.6. Performance measures for experiment one	156
Table 5.7. Mobile phone call locations for experiment one	158
Table 5.8. Descriptive data for the individual difference measures for experiment one	158
Table 5.9. Score range and means for the performance and attitude measures for experiment one	159
Table 5.10. Main effects for subjective measures for experiment one	160
Table 5.11. Main effects for objective measures for experiment one	164

Table 6.1. Interface objects, menu options, and dialogue for the 2 services used for experiment two.....176

Table 6.2. Percentage of calls for both private and public locations for experiment two.....177

Table 6.3. Performance measures for experiment two179

Table 6.4. Attitude measures for experiment two179

Table 6.5. Descriptive data for the individual difference measures for experiment two183

Table 6.6. Score range and means for the performance and attitude measures for experiment two.....184

Table 6.7. Descriptive statistics for performance measures across the 6 trials184

Table 6.8. Descriptive statistics for attitude measures across the 6 trials187

Table 6.9. Descriptive statistics for performance measures for private and public use188

Table 6.10. Descriptive statistics for attitude measures for private and public use ..189

Table 6.11. Predictors of private performance for the standard service.....192

Table 6.12. Predictors of public performance for the standard service.....192

Table 6.13. Predictors of private performance for the office filing system service ..193

Table 6.14. Predictors of public performance for the office filing system service ...193

Table 6.15. Predictors of private attitude for the standard service195

Table 6.16. Predictors of public attitude for the standard service	195
Table 6.17. Predictors of private attitude for the office filing system service	196
Table 6.18. Predictors of public attitude for the office filing system service	196
Table 6.19. Summary of significant predictors of performance and attitude	197
Table 7.1. Performance measures for experiment three	207
Table 7.2. Attitude measures for experiment three	208
Table 7.3. Descriptive data for the individual difference measures for experiment three.....	210
Table 7.4. Score range and means for the performance and attitude measures for experiment three.....	210
Table 7.5. Descriptive data for the performance measures for the 3 experimental tasks.....	211
Table 7.6. Main effects between groups for the attitude measures for experiment three.....	212
Table 7.7. Main effects between visualisers and non-visualisers for the performance measures for experiment three	214
Table 7.8. Main effects between visualisers and non-visualisers for attitude measures for experiment three	214
Table 7.9. Question 1: What aspects of the service did you visualise?.....	217
Table 7.10. Question 2: What strategy did you use for navigating through the service?.....	218

Table 7.11. Question 3: What features of the service helped you to know where you were?	219
Table 7.12. Question 4: How well did you remember the structure of the service from task to task?	220
Table 7.13. Question 5: How did you feel using the service for the first time?.....	221
Table 7.14. Question 6: At what point did you start to feel confident using the service?.....	222
Table 7.15. Question 7: How does this service compare to the standard menu style service?.....	223
Table 7.16. Design factors for automated mobile phone services.....	224
Table 7.17. Interview quotes from the 3 services supporting the 18 design factors	225
Table A11.1. Variable labels assigned to the Technographic questionnaire items ...	311
Table A11.2. Rotated component matrix from the PCA.....	312
Table A11.3. The 5 components produced by the principal components analysis ...	312
Table A11.4. Percentage variance for each component after rotation	313
Table A11.5. Descriptive data for the PCA components	315
Table A21.1. Absolute times for preliminary study two	331
Table A21.2. Absolute times for experiment one.....	331
Table A21.3. Absolute times for experiment two	331

Table A21.4. Absolute times for experiment three331

ACKNOWLEDGEMENTS

Many people helped and supported me in many different ways during the course of my thesis. In particular, I would like to thank my two supervisors: Dr Steve Love of the Department of Information Systems and Computing at Brunel University, and, Dr Mark Turner of the Department of Psychology at the University of Portsmouth. I am grateful for their invaluable advice, their encouragement when times were tough, their friendship, and their support throughout. Thanks also to Dr Darren Van Laar from the Department of Psychology at the University of Portsmouth, who acted as a third supervisor during the first year of the PhD, and who contributed greatly during the formative stages of this research work.

Thank you to Paul Waby from the technical services unit at the Department of Psychology at the University of Portsmouth, who provided excellent support in setting up apparatus, and providing materials for the early studies conducted for this PhD. Thanks also to John Park from the technical support unit at the Department of Information Systems and Computing at Brunel University, who generously provided experimental equipment, and software licences, which significantly enhanced the quality of the experiments conducted. A big thank you to all of the students, colleagues, friends, and family who participated in the studies and experiments conducted for this thesis, without whom the empirical work, and therefore the thesis, would not have been possible.

Special thanks must go to my partner Marie Norberg, who has been a source of continual encouragement, and has put up with me being 'lost in thought', sometimes for weeks at a time. Thanks also to my friends, who have provided a welcome distraction from the PhD when it was needed most. Finally, I would like to thank my mother, Carole, who has always encouraged me in my academic pursuits, given me the self-belief to take on new challenges, and supported me wholeheartedly.

I dedicate this thesis to the memory of my father, Michael Howell.

DECLARATION

The following papers have been published, or submitted for publication, as a result of the research conducted for this thesis.

HOWELL, M.D., LOVE, S., and TURNER, M., 2005, Spatial metaphors for a speech-based mobile city guide service. *Journal of Personal and Ubiquitous Computing*, **9**, 32-45.

HOWELL, M.D., LOVE, S., and TURNER, M., 2005, The impact of interface metaphor and context of use on the usability of a speech-based mobile city guide service. *Behaviour and Information Technology*, **24**(1), 67-78.

HOWELL, M.D., LOVE, S., and TURNER, M., (in submission), Visualisation improves the usability of voice-operated mobile phone services, *International Journal of Human-Computer Studies*.

HOWELL, M.D., LOVE, S., TURNER, M., and VAN LAAR, D.L., 2003, Generating interface metaphors: a comparison of 2 methodologies, In: McCabe, P.T. (ed.) *Contemporary Ergonomics 2003* (London, UK: Taylor & Francis), pp. 235-240. ISBN 0-415-30994-8.

HOWELL, M.D., LOVE, S., TURNER, M., and VAN LAAR, D.L., 2003, Interface metaphors for automated mobile phone services. *Proceedings of HCI International 2003, 10th International Conference on Human-Computer Interaction, Crete, Greece, June 2003*, pp. 128-132. ISBN 0-8058-4930-0.

HOWELL, M.D., LOVE, S., TURNER, M., and VAN LAAR, D.L., 2004, Improving the usability of mobile phone services using spatial interface metaphors. In: McCabe, P.T. (ed.) *Contemporary Ergonomics 2004* (Florida, USA: CRC Press), pp. 235-240. ISBN 0-8493-2342-8.

:: CHAPTER 1

Introduction

1.1 Introduction

Since the design of the desktop metaphor interface for the 8010 Star Information System (Smith, Irby, Kimball, Verplank, and Harslem, 1982) the use of metaphor has been recommended as an important technique for graphical user interface (GUI) design: ‘A physical metaphor can simplify and clarify a system’ (Smith et al., 1982, p. 655); ‘The use of interface metaphors has dramatically impacted actual user interface design practice’ (Carroll, Mack, and Kellogg, 1988, p. 67); ‘Metaphors have two distinct but related uses in interface design: as cognitive aids to users, and as aids to creativity for designers. Metaphors can help designers to use their own, often unconscious, expectations to create new information links and mental structures’ (Mountford, 1995, p. 137); ‘Understanding and improving strategies for developing user interface metaphors is one of the most important and formidable goals for human-computer interaction’ (Neale and Carroll, 1997, p. 441); ‘Interface metaphors have proven to be highly successful, providing users with a familiar orienting device and helping them understand and learn how to use a system’ (Preece, Rogers, and Sharp, 2002, p. 56); ‘Tremendous commercial successes in computing have arisen

directly from a judicious choice of metaphor...Very few will debate the value of a good metaphor for increasing the initial familiarity between user and computer application' (Dix, Finlay, Abowd, and Beale, 2004, p. 169).

Despite these claims about the advantages of metaphor, the use of interface metaphor has largely been limited to GUIs. The primary aim of the work reported in this thesis is to investigate whether interface metaphor can be successfully used to implement speech-based automated mobile phone services, with success being judged in terms of an improvement in usability compared to currently existing automated mobile phone service designs.

This chapter begins by discussing the motivation behind the research programme conducted, and presents a brief background to the research. These discussions form the basis for identifying the three primary research questions. The research methodologies employed to investigate the research questions are then introduced. Finally, an outline of the thesis structure is presented, giving a brief description of the contents of the remaining chapters.

1.2 Research motivation

Mobile phones are widespread, with the number of mobile subscribers estimated to be more than 1 billion globally (PC Advisor, 2003). Growth in the availability and use of mobile phones has created both demands, and opportunities, for the design of new services that are quick to learn, and easy to use. These services can be accessed at any time from any place; a scenario that raises two major usability issues. Firstly, mobile phones typically have small screens, which limits the amount of information that can be visually displayed. Secondly, if services are to be accessed in both private and public locations, both indoors and outdoors, then the effects of context of use on their usability must be investigated.

Attempts have been made simply to apply the same graphical user interface (GUI) techniques from the desktop computer to the mobile phone, but the resulting interfaces may be difficult to use for a number of reasons. Jones, Marsden, Mohd-Nasir, Boone, and Buchanan (1999) found that small screens have a negative effect on performance for both focussed and less directed search tasks for web-based data.

Dillon, Richardson, and McKnight (1999) found that when using small screens, the use of the page backwards and page forwards options increased, whilst Han and Kwahk (1994) showed that search times on single line displays were far slower than on desktop displays. In addition to the limitations of small screens, current data input techniques for mobile phones, such as keypad input, can be slow and tedious, whilst mobile Internet browsing has been limited to purpose built sites which often suffer from slow connection speeds. These problems may be amplified when the mobile phone is used in busy public places, as the environment competes with the interface for the user's perceptual resources (Kjeldskov 2002).

An alternative to keypad input is speech, and an alternative to using the Internet for information access is the use of automated phone services, which currently provide speech output, and require users to navigate and select items using either speech, or the keypad. Speech is a natural and familiar form of human communication, which suggests that these services should be easy to use. However, there are problems with these services, which involve the user getting lost in the menu structure (Wolf, Koved, and Kunzinger, 1995), and not being able to remember the menu options (Schumacher, Hardzinski, and Schwartz, 1995). It is possible that the use of an appropriate interface metaphor, which has been a successful design technique for GUIs, may alleviate some of these usability problems by representing the structure of the service in terms of a familiar structure from the real world, thereby allowing users to transfer their existing knowledge to the service. Investigating the use of interface metaphor as a means of improving the usability of speech-based mobile phone services provides the central motivation for this research.

There are different factors that affect a person's interaction with a computer system, and a model of HCI proposed by Eason (1991) demonstrates the relationship between these factors (see Figure 1.1). Within the model, human-computer interaction is considered as being analogous to a conversation between a 'human' party and a 'computer' party. The goal of the conversation is to perform a 'task', but the 'environment' can affect task performance. Each of the components interacts with the others, and all of the components contribute to the design requirements, and must therefore be considered as part of the design process.

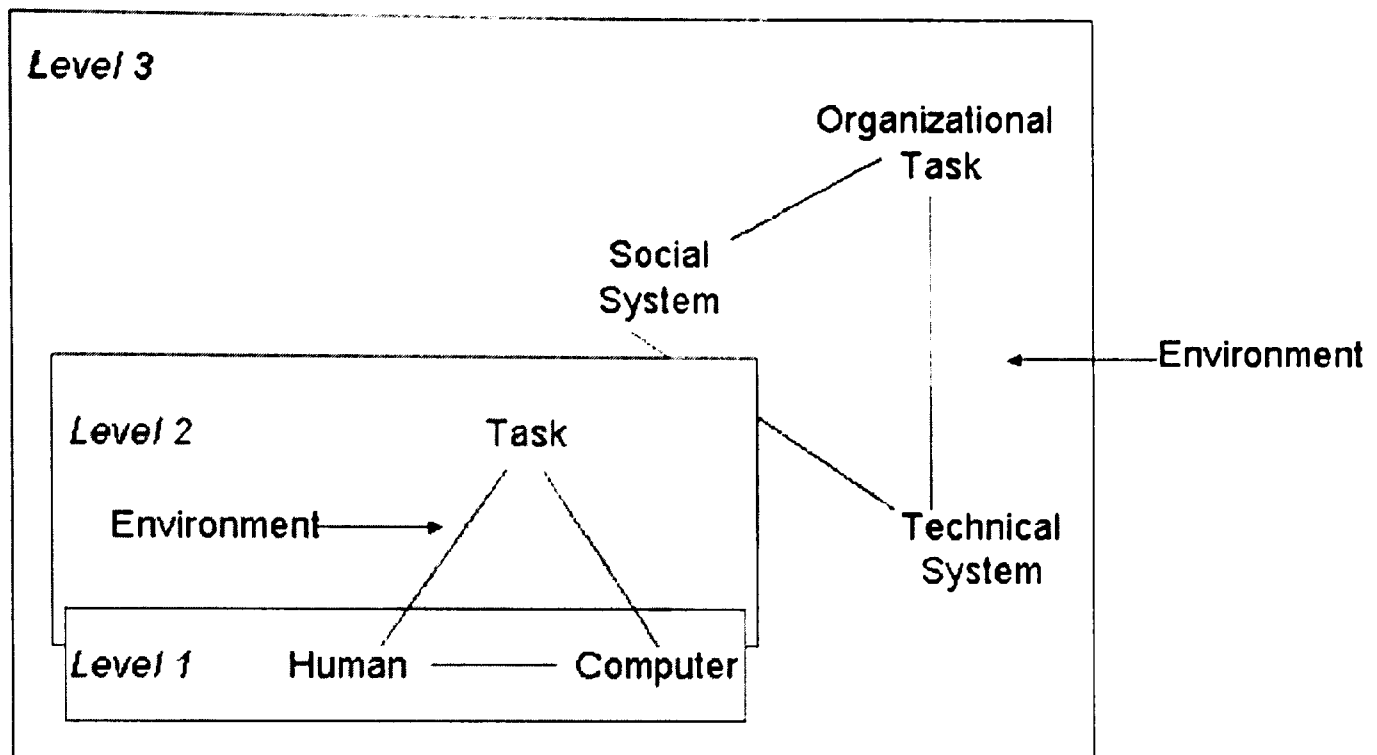


Figure 1.1. The three-level model of HCI (Eason, 1991)

Level one concerns the user interacting with the computer, with both participants in this interaction being capable of information processing. Level two broadens the framework to consider the task and the environment in which the tasks are performed. The environment may also be referred to as the context of use and refers to the physical, organisational, and social contexts of use. The third level considers the factors that are important when HCI takes place within a much broader setting, which has an impact on social life, the way organisations operate, and the way individuals behave.

For the research work conducted in this thesis, the interaction between the components in levels one and two will be considered. The ‘computer’ component of the model refers to speech-based automated mobile phone services, and the use of interface metaphor as a technique for designing these services. The ‘human’ component refers to the users’ individual differences, which are the characteristics, and aptitudes that may impact their performance with, and preferences for, these services. The ‘environment’ component of the model refers to the physical and social contexts in which these services are used. The ‘task’ component of the model refers to the tasks that users want to perform. By attempting to perform tasks with an automated mobile phone service, the effects of interface metaphor, individual differences, and context of use on task performance may be examined. The task component will therefore be treated as a catalyst for the investigation of the other

three components. The three main research questions driving the research programme reported in this thesis are each related to one of the three components from the Eason (1991) model (computer, environment, human), and can be seen below:

- Question 1: Can interface metaphors improve the usability of speech-activated automated mobile phone services?
- Question 2: Does context of use affect the usability of metaphor-based speech-activated automated mobile phone services?
- Question 3: To what extent do the individual characteristics of users affect the usability of metaphor-based speech-activated automated mobile phone services?

This section has outlined the research motivation, and presented the three main research questions, which will be transformed into research objectives after the relevant literature has been reviewed in chapter two. The following section will describe the experimental methodologies that will be employed for the research work reported in this thesis.

1.3 Research methods

The research work reported in this thesis was conducted in two stages. The first stage involved the design, implementation, and testing of prototype mobile phone services, and the second stage involved the comparison of these service prototypes. For stage one, a Human Centred Design (HCD) methodology was followed in order to generate and develop metaphors for the design of the metaphor-based versions of the mobile phone service. HCD ensures that an interactive product is designed with high levels of usability. The three original principles of HCD (Gould, Boies, Levy, Richards, and Schoonard, 1987) are: to involve users as early, and as much as possible during the design process so that users' cognitive, social and attitudinal characteristics are understood and accommodated; to measure performance and attitude by using interfaces and simulations of the system; and to design iteratively so that testing and evaluation can be conducted to check that the design meets the user requirements.

Stage two involved a comparison of the usability of different prototype services under different conditions in order to address the research questions. An experimental methodology was employed to provide a controlled framework for the manipulation of the independent variables of interest, and to measure their effect on the dependent variables. Field experiments, rather than laboratory experiments, were conducted in order to more accurately reflect the everyday conditions in which mobile phone services are used, and to subsequently increase the ecological validity of the results. Chapter three provides a detailed discussion of the methodologies used.

1.4 Thesis overview

Chapter two considers previous research that has been conducted within HCI in four main areas, and has subsequently been divided into four main sections. The first section covers automated telephone service design issues, and is followed by a section that presents a review of interface metaphor. The third section introduces the psychological study of individual differences within HCI, and the final section addresses the issue of context of use.

Chapter three is divided into two sections, corresponding to the two methodologies used for the work conducted for this thesis. A description and justification is provided for the selection, and use of, both the HCD methodology, and the experimental methodology. For the HCD process, the five design stages are described, including the full range of techniques and materials used, with a detailed account of the Wizard of Oz (WOZ) technique (Fraser and Gilbert, 1991) used for the prototype simulations. For the experimental methodology, operational definitions for the variables are stated, followed by a general discussion of the experimental design, participant sample, and the range of materials and apparatus used for the three main experiments.

Chapter four is the first experimental chapter, and describes two preliminary laboratory-based studies conducted. The first study aimed to explore relevant methodologies for the generation, selection, and development of interface metaphors, leading to a number of metaphor categories that may be applicable to automated mobile phone services. Preliminary study two was a user testing study that was conducted to evaluate the metaphor-based phone service prototypes, and to provide feedback for their redesign.

Chapter five describes the first field experiment conducted, which compared the usability of three metaphor-based versions of a mobile city guide service with the usability of a non-metaphor version of the service. The experimental methodology is presented, including the design used, the variables measured, the participant sample, the apparatus, the procedure followed, and the data collection instruments utilised. The results are then documented, followed by a discussion section, and concluding with a chapter summary.

Chapter six describes the second field experiment, designed to investigate the usability of a metaphor-based mobile city guide service over an extended period of time, and to investigate the effect of both private and public context of use on user's interaction with the service. A range of participants' individual differences were also measured and analysed to explore the association between individual differences and the usability of the services. As in chapter five, the methodology, results, and discussion are presented, concluding with a chapter summary.

Chapter seven presents the final field experiment, which was designed to qualitatively investigate the visualisation strategy used by some participants when interacting with the metaphor-based version of the mobile city guide service. In addition, the experiment also sought to assess whether a computer-based metaphor could be successfully applied to the mobile city guide service. The experimental methodology and results are documented, followed by a discussion, and concluding with a chapter summary.

Chapter eight presents a summary of the research findings from chapters four, five, six, and seven. A discussion is then presented of the findings of this thesis, and the contribution these make to the understanding of the ways in which, and the extent to which, spatial interface metaphors may be used to improve the usability of speech-based automated mobile phone services. The chapter also identifies potential limitations of the research work conducted, and possible areas for future experimental research that may extend the current research findings.

:: CHAPTER 2

Literature review

2.1 Introduction

The main aim of the research work reported in this thesis is to assess the usefulness of interface metaphor as a technique for the design of speech-based automated mobile phone services. This chapter provides a review of relevant background literature, which is structured around the model of HCI proposed by Eason (1991), which was discussed in chapter one. Automated phone services, interface metaphor, individual differences, and context of use will therefore be reviewed within this chapter. This chapter will demonstrate that the use of interface metaphor may be an effective technique for designing speech-based mobile phone services. It will also establish that, in order to assess this effectiveness, both the users' individual differences, and the contexts in which the phone services are used, must be considered.

This chapter begins by examining current knowledge on the design of automated phone services, followed by a discussion of conversational interfaces and the problems involved with implementing them. The usability problems of speech-based mobile phone services are then covered, followed by a review of three different

approaches that have been adopted to overcome these problems. Section 2.3 discusses the use of interface metaphor within HCI, highlighting the lack of research investigating the application of metaphor to speech-based systems. This section includes sub-sections covering the problems of using metaphor, alternatives to metaphor, categories of metaphor, metaphors for different computer systems, and guidelines and frameworks for integrating metaphor into the design process. Section 2.4 provides a review of the psychology of individual differences within HCI, and a discussion of the individual differences that are considered to be of most relevance to interaction with speech-based mobile phone services. Section 2.5 discusses context of use, and the reasons for its importance as a factor affecting interaction with speech-based mobile phone services. Physical and social context of use are shown to be most salient to the mobile phone-based research work reported in this thesis. Finally, in section 2.6, conclusions from the literature review will be presented, in order to establish a clear rationale for the research work reported in this thesis, and to highlight the importance of each of the three components of HCI examined: computer, human, and the environment. The three primary research objectives driving this programme of research are also presented.

2.2 Automated phone services

2.2.1 Introduction to automated phone services

Automated telephone services are becoming increasingly pervasive within the telephone network (Stentiford and Popay, 1999), allowing users to access services such as voice-mail, film schedules, transport timetables, and personal banking. These services currently use speech as output, and offer the user a choice of speech and/or keypad for input, for example, British Telecoms CallMinder telephone answering service (Beacham and Barrington, 1996). These services are usually structured hierarchically, and consist of lists of spoken menu options assigned to numbers, which, when selected, lead to sub-menus (Hallstead-Nussloch, 1989). In order to progress forwards and backwards through the service, the user selects a menu option by remembering the number corresponding to the desired option, and then either saying the number, or pressing the corresponding keypad key (Schmandt, 1994). One of the key benefits of using voice menus is that they require little training, allowing

the novice user to access the system and effectively perform tasks (Marics and Engelbeck, 1997).

The growth of such services has been driven by their ability to utilise speech technology to allow information to be accessed over the telephone, 24 hours a day, seven days a week (Rabiner, 1995; Strathmeyer, 1990). This increases flexibility and convenience, and reduces operating costs by replacing human operators (Whittaker and Attwater, 1996). It has also been claimed that voice control makes these services easier to use and therefore improves customer satisfaction (Westall, Johnson, Lewis, 1996), but this claim is not supported by empirical evidence. A more obvious benefit of speech control is in extending system usage to areas where the user is performing eyes-busy and/or hands-busy tasks (Martin, 1989). Previous systems that have capitalised on this have been applications for quality control and inspections, stock control, parcel sorting (Visick, Johnson & Long, 1984), baggage handling (Nye, 1982; Jones, Frankish & Hapeshi, 1992), meter reading (Markowitz, 1993), office systems (Noyes and Frankish, 1989), military applications (Chambers and deHaan, 1985; Moore, 1989) and direct speech input to computers for medical and dental procedures (Martin, 1976). Speech can also provide a means of access for people who find difficulty in using computers (Noyes, Haigh, and Starr, 1989).

Within telecommunications, automated speech services have been designed for a range of application areas such as voice messaging, banking, betting, information services such as transport timetables, telemarketing, telephone shopping, remote access to electronic mail, and automated operator services for call centres (Westall et al., 1996). The proliferation in the use of mobile phones means that people can now access these services at any time and from any place. In order to ensure the usability of such services, Schmandt (1994) points out that ‘the awareness that speech takes time (and acceptance that time is a commodity of which we never have enough) should permeate the design process for building any speech application.’ (Schmandt, 1994, p. 106). To achieve this, Hallstead-Nussloch (1989) suggests that designers consider both the mode of communication, and the dialogue design. These two aspects will be considered in the following two sub-sections. Section 2.2.1.1 assesses the benefits and drawbacks of using speech as a mode of human-computer

communication, and section 2.2.1.2 reviews the existing guidelines for the dialogue design of automated phone services.

2.2.1.1 Mode of communication

Mode of communication refers to the means of input available to the user, and the corresponding output provided by the system as a response. In terms of input, designers must choose either speech input, keypad input, or both. Hallstead-Nussloch (1989) concedes that voice input is more intuitive, but highlights the problems associated with accurate recognition of this speech by the system. He recommends speech input as being a more convenient mode of interaction when the user's eyes or hands need to be free to perform some other task, when a small number of system commands is used, and when the system is used in a quiet environment where the cost of making an error is low. One of the main advantages of speech input is that the operator can be free of the constraints of visual displays (Wilpon and Roberts, 1986; Rollins, Constantine, Baker, 1983). The time taken to examine a visual display for the correct icon, menu item, or number can introduce considerable delay into the task cycle (Jones, Miles and Page, 1989), whereas a speech input system designed with a limited spoken vocabulary and a well-structured grammar offers a faster and more direct style of interaction (Martin, 1989).

The advantages of speech output also centre on its effectiveness in supporting users at times when focussing on a visual display may be inconvenient, for example, when they are mobile, or concentrating on written material (Schmandt, 1994). In fact, empirical studies have shown that when a user is performing multiple tasks, their performance improves if the tasks can be managed over separate input/output modalities (Allport, Antonis and Reynolds, 1972; Wickens, Mountford, and Schreiner 1981; Treisman and Davies, 1973). Jones, Hapeshi and Frankish (1989) also discovered that a speech-based task such as operating an automated phone service, may be effectively combined with other tasks, but only if they are non-verbal. When two activities of different types are undertaken, for example a verbal and spatial activity, they do not interfere with each other as strongly as they would if they were of the same modality (Baddeley and Hitch, 1974). This suggests that speech interaction with a mobile phone service could be effectively combined with the spatial navigation of a person's everyday environment without a significant performance detriment,

although background speech may prove to be disruptive to memory and information processing.

Another advantage of speech output concerns humans' improved ability to recall auditory information over visual information for short retention periods (Engle, 1974, Crowder, 1970; Murdock and Walker, 1969). This effect may be attributed to 'echoic' memory (Crowder, 1976; Neisser, 1967) which serves to retain information for a short period of time, even when the person is not consciously paying attention to the sound source at the time of presentation. The visual equivalent is known as iconic memory (Neisser, 1967), but iconic memory for a visual medium such as text tends to degrade more quickly than the echoic memory for speech (Wickens, 1992). Speech output for the short serial lists of menu options that characterise automated phone services may therefore lead to improved recall when compared to visual presentation (Hapeshi, 1993).

These advantages of speech control are well suited to accessing automated phone services from a mobile phone, when the user may be dividing their attention between other simultaneous tasks such as walking along a pavement or crossing a road. In such situations, speech interaction would be more convenient for the user as it both frees up the eyes for other tasks, and removes the need to take the handset from the ear in order to input using the keypad.

The problem with using speech output is that speech is slow, serial, and temporal, which means that listening to spoken information can become tedious, and requires the user to either remember what was said, or have to request a repeat of the whole system message (Schmandt, 1994). In addition to this, privacy issues exist, as voice announcements can be heard by other people, whereas with visual feedback it is easier to keep the information private.

A further negative side-effect of using speech output is caused by users projecting human-like qualities onto the speaker. When a system is designed to sound too human-like, some users anthropomorphise the system, and in so doing endow it with qualities far beyond its capabilities, such as intelligence and power. This leads users to develop unrealistic expectations of the system, which can lead to inappropriate

behaviour (Hapeshi and Jones, 1988), which highlights the need to match the users needs and expectations with the functional capabilities of the system. Studies have also shown that there is a tendency for synthetic speech to be perceived as being ‘unfriendly’, ‘harsh’, or ‘evil and sinister’ (Michaelis and Wiggins, 1982; Edman and Metz, 1983; Love, Foster, and Jack, 2000), and Peacock (1984) notes that some users just do not like being told what to do by speech systems.

2.2.1.2 Dialogue design

Operating automated phone services involves a process of explicit user selection followed by a system response, which requires that the dialogue designer assume responsibility for both the user’s and the system’s role in the interaction. This can be contrasted with human-human interaction where both parties have an implicit role and responsibility to maintain the dialogue flow, which is learned over time. Dialogue flow is the movement from the system’s role in the dialogue to the user’s role, is ongoing until completion of the interaction, and is a critical factor in the success of an automated phone service (Aucella and Ehrlich, 1986; Aucella, Kinkead, Schmandt, Wichansky, 1987).

Command-based flow is where the user generates commands without being prompted by the system, and is suitable for systems that have a narrow range of system functionality that is used frequently (Hallstead-Nussloch, 1989). An example of a system using command-based flow is the Speech Filing System (SFS) designed by Gould and Boies (1983). The SFS is a voice store and forward office system that allows the user to create a voice message by calling a designated telephone number, and then to listen to and edit their message by pressing predefined telephone keypad keys. The SFS has four modes corresponding to four categories of functionality: record, transmit, get, and listen. Within each mode the user is allowed to perform a number of actions, the first letter of which corresponds to the letter on the keypad key. For example, to enter the ‘transmit’ mode the user presses the ‘T’ key, and once in this mode the user can send a message to a person’s ‘name’ by pressing the ‘N’ key. Problems with command-based flow include the need for training to learn the command keys, and the burden on working memory caused by having to remember these commands. In fact, in order to counter these problems, the SFS system gradually evolved from being a system that demanded user initiated commands to being one

that prompted users with spoken menus. This prompted style of interface has become the most widely used for implementing speech systems.

Prompted flow is where the system provides the user with a number of choices or actions, and the user responds to these (Hallstead-Nussloch, 1989). Prompted flow is suitable for more extensive systems, with a larger number of commands, and which are used infrequently. In designing the system's role in a prompted style of service, the designer must consider two dialogue components (Hallstead-Nussloch, 1989), the system prompt and the system message.

The system prompt is the part of the dialogue where the user is given menu options to choose from in order to progress through the service. These menu options may be either temporal or enumerated. Using the example of a city information service, an enumerated menu would assign a number to each menu option, for example, 'For Brighton, say 1; for Liverpool, say 2; and for London, say 3'. A temporal menu would ask the user to press a key when they hear the choice desired, for example, the dialogue might resemble the following, 'Hit any key when you hear the city you want. Brighton...Liverpool...London.'

An advantage of the number-based menu style is that it is easy to learn for novice users, whilst allowing experienced users to interrupt the prompts without having to listen to the whole dialogue (Schmandt, 1994). Temporal menus tend to be slower, requiring the user to wait until the desired choice is presented, which may suit novice users, but could be frustrating for more experienced users. This highlights the fact that the skill of a speech-interface user varies with experience and practice (Leggett and Williams, 1984), with a novice user having very different abilities to those of the experienced user (Norcio and Stanley, 1989). For example, inexperienced users prefer a rigid prompted interaction style (Zoltan-Ford, 1991) that reduces the chance of making mistakes (Morrison, Green, Shaw and Payne, 1984), and provides them with timely feedback on their actions and commands (Poock, Martisa and Roland, 1983). In contrast, an experienced user may find such an interface too rigid, long-winded, boring, poorly focused, ineffective and sometimes misleading (Brajnik, Guida and Tasso, 1990).

Hallstead-Nussloch (1989) claims that the effectiveness of auditory menus is linked to their ability to present all of the relevant information to the user, and that in order for this to happen they must contain three components. The first component is a title, which informs the user of their position within the service. The title helps prevent the user from becoming lost, and primes them for the forthcoming menu options. The second component is a list of menu options, which should be presented in a 'goal-action' sequence, for example, 'to select this option, press 1', as this is more consistent with the cognitive make up of the task, and reduces the short term memory load. On the basis of their experience implementing a prototype voicemail system for athletes at the 1984 Olympics, Gould et al. (1987) found that four menu items was too many. However, later work by Engelbeck and Roberts (1989), and Bond and Camack (1999), confirmed four menu items as being a comfortable number, but that larger numbers of options may be used for more experienced users, as these users demonstrate greater confidence, and interrupt the dialogue more often. The third component is the ending, which informs the user that their interaction with the menu is complete.

Marics and Engelbeck (1997) suggest that the ability to interrupt prompts is critical, allowing more experienced users to interrupt dialogue irrelevant to their task, allowing them to progress more quickly through the service. They also recommend a 'type-ahead' function which allows those users who remember the service structure and options to interact more efficiently by entering the numbered pathway sequence without waiting for the dialogue prompts between their start and end point. A final recommendation is the provision of 'time outs', which come into effect if the user has not responded to the system prompts within a specified period of time. This may be due to a number of factors, for instance, they may not have understood the prompt, they may not know what to enter, or they may be lost within the service and do not know how to exit. In such cases, after a period of 5 to 10 seconds the system should repeat the prompt or offer additional help, eventually transferring the user to a human operator after three time out events (Marics and Engelbeck, 1997).

The system message is the part of the dialogue that informs the user of their current status and location within the service, and helps to maintain the flow of the conversation. Hallstead-Nussloch (1989) observed that three different types of

messages can be used within automated phone services, but do not necessarily prescribe the use of all three: error messages inform the user of a user or system generated error, and how to recover from it; completion messages provide the user with feedback on their most recent command; and working messages inform the user that a particular command is being processed, and may provide a time estimate for completion. Schumacher (1992) suggests carefully wording the opening message of a service, so that the user knows which service they have reached. He also proposes that users be told how many menu items to expect as a way of reducing erroneous input selections caused by user uncertainty, for example, 'There are 3 categories of restaurants available, for Indian restaurants, say 1....'.

Whilst interacting with automated phone services, user's activities are limited to (1) listening to system prompts and messages (2) selecting menu options. In addition to selecting menu options, the user should be provided with additional control functions (Halstead-Nussloch, 1989), enabling them to navigate through and control the dialogue presentation. As well as a prompt interrupt feature, he also suggests: allowing users to repeat the prompt or message they have just heard; providing a help facility if they become lost or unsure of what to do next; and finally a global command to return the user to the beginning of the service.

Schumacher et al. (1995) propose that most conversational pleasantries such as 'please' and 'thankyou' are not appropriate to phone services, and that the inclusion of such dialogue can lead to both inefficiency and confusion, as user expectations of a machine-like interaction are temporarily breached. They state that the goal for designers should be a balance between efficient dialogue that is not perceived as abrupt or rude, and advise that prompts should be worded using simple, explicit, concise terminology that is related to the task domain.

Choinere, Robert and Descout (1991) recommend providing an adaptive help facility that evolves through the course of the interaction to offer the user help that is specific to their current location within the service. Devauchelle (1991) recommends having all prompts and messages recorded by the same speaker, regardless of whether the voice is synthetic or real. Schmandt (1994) reinforced this point with his finding that when different voices are used, a primary voice usually emerges as being preferable to

and more intelligible than the other voices. However, preferences for voice types are dependent on whether the task involves information access, entertainment or feedback (Rosson and Cecala, 1986) and must be evaluated accordingly.

For speech recognition systems, Shneiderman (1992) suggests that in order to narrow the size of the speech-interface vocabulary, the application and the domain must be considered in the design. Kamm and Helander (1997) state that, from the user's perspective, the vocabulary must be memorable and meaningfully associated with the task, whilst from the designer's perspective the vocabulary range must be limited, and selected to ensure that the recogniser does not confuse words. To ensure memorability they recommend a user-centred approach to selecting the vocabulary, whereby users choose words that they associate with the task. They also suggest the use of synonyms to increase the flexibility of the system, although the range of synonyms must be limited to those within the vocabulary of the speech recogniser.

The previous two sections have discussed the advantages and disadvantages of speech interaction with computers, and reviewed existing guidelines for the design of prompted style automated phone services. The following section considers the differences between human-human communication, and human-computer communication, and the difficulties these present for the speech interface designer.

2.2.2 Conversational automated phone services

Enumerated prompted style interfaces were not originally designed according to a human-human analog, but were originally designed to be accessed from a keypad (Kamm and Helander, 1997). An automated telephone banking service requiring keypad input may have used the following dialogue 'for checking account balance, press 1; for money transfer, press 2'. Early speech recognition versions of these services simply translated the menu options from keypad to voice command, for example, 'for checking account balance, say 1; for money transfer, say 2'. Voice-activated versions of these services exploit the benefits of a dual combination of auditory and visual interface elements (Vetere and Howard 1999) by utilising the visual cue of the number on the phone itself and the auditory 'label' assigned by the spoken dialogue. However, speech recognition systems are not tied to the numerical keypad-based menu options, and are able to utilise a more meaningful vocabulary.

which should result in a more conversational style of interaction. A more conversational style of interaction for the telephone banking service may have asked a question such as, 'Would you like to check your account balance or transfer money?'

Wolf, Kassler, Zadrozny and Opyrchal, (1997) suggest a number of advantages to a conversational style of interaction. Firstly, the user is able to formulate requests in their own words, rather than selecting a menu option that best matches their goal. Secondly, a conversational interface allows the user to achieve their goals directly without having to navigate through a hierarchical menu structure. Finally, users can request more than one piece of information at a time, rather than having to break the overall goal down into a number of discrete tasks. The speech-recognition system would then be required to separate out a multiple request for information into its component sub-tasks, execute each of the sub-tasks, and finally provide the information to the user in a logical sequence. Conversational interaction promises a less structured transaction, where the user can explicitly state their goals rather than having to navigate through a hierarchy of menu options.

Yankelovich (1994) proposed the following guidelines for the design of conversational interfaces (1) maintaining the flow of conversation through the use of barge-in technology (the same feature as type-ahead for keypad input) which allows the user to respond at any time. Mixed initiative dialogue should also be a feature enabling either the user or the system to take the conversational lead (2) maintaining a shared situational context, leading to an awareness by both parties of the perceptual and conceptual aspects of the situation in which the conversation is occurring (3) providing sufficient feedback to enable the user to recover from recognition errors made by the system.

Peckham (1993) provides additional guidelines, which derive from the often unstructured and spontaneous use of speech by humans. He believes that human-computer conversations should occur in real time, as they do in human conversation. He proposes that conversational interfaces allow for spontaneity and extended interaction, allowing a series of turn-taking events to occur, as is the norm for natural dialogues. Finally, he emphasises the need for a cooperative element to the interaction, so that if the system fails to locate the information requested by the user,

an explanation is provided rather than simply a negative response. Despite the proposed advantages of a conversational style of interaction, and the existence of guidelines for the design of conversational interfaces, conversational automated phone services are not commercially available. Rather, prompted style automated phone services are still dominant. The following two sub-sections will examine the two main reasons for this, which are proposed as being the differences between human-human communication and human-computer communication, and the difficulty in achieving accurate speech recognition, over a telephone connection, for the large vocabulary required by a conversational interface.

2.2.2.1 Components of human conversation

Humans have been talking to each other using some form of speech for the last 50 000 years (Westall et al., 1996), and as speech is a natural communication medium between humans, it is often assumed that it would also be a more natural communication medium for human-computer interaction than other methods such as keyboard or mouse input (Hapeshi, 1993). Some researchers have even assumed that due to the 'naturalness' of speech, interfaces should be speech-based whenever possible (e.g. Lea, 1980). However, in order for the use of speech to be considered natural, the interaction between the human and the computer must also be natural, suggesting that it should adhere to human conversation principles.

Newell (1992) highlighted the fact that speech researchers had become fixated on the technological barriers to widespread adoption of speech recognition systems, such as the intelligibility of speech synthesis and speech recognition accuracy, whereas more emphasis should have been given to designing the dialogue for effective and efficient interfaces. He suggested a shift in focus from the technological problems, to the problems caused by attempting to map human-human interaction onto human-computer interaction. Previously, an underlying assumption had been that the skills and expectations gained from the everyday conversational use of speech could be effectively transferred to human-computer interaction. It was now evident that the two styles of interaction were fundamentally different. As a result, Damper (1993) emphasised the need for a thorough understanding of the unique nature of human-human speech communication, and a realisation that speech is not a universal panacea for all interfaces, but that it may be well suited to specific circumstances.

To investigate the differences between the two types of interaction, a definition of speech interaction offered by Baber (1993) will be considered as a useful starting point: 'speech interaction may be defined as a process by which meaningful information is communicated between parties, using a sequence of speakers alternating turns.' (Baber, 1993, p. 4). Therefore, a prerequisite for speech interaction is some understanding of the language being used by both parties, which may be hard to attribute to a computer, and also a sequence of speakers, without which a monologue would occur. The 'sequence of speakers' is managed by 'turn-taking' strategies (Sacks, Schlegoff, and Jefferson, 1974), which draw on a wide range of linguistic and extralinguistic cues not available to human-computer communication (Duncan, 1972). As a result Karis and Dobroth (1991) argue that the rhythm of human conversation is never really obtained in human-computer dialogues.

Nickerson (1976) points to a number of unique characteristics of human conversation which are difficult to transfer to human-computer conversation. The first of these is mixed initiative, which refers to the way in which one party in a conversation is allowed to direct the interaction, often unprompted or in the form of a question, but always in a way that makes it clear at all times who is currently in control. This is often achieved through the use of non-verbal cues. In human-computer conversation, the implementation of an interrupt function may provide the user with the opportunity to take the initiative, but this is only a very rudimentary form of mixed initiative, and lacks the richness and variety of the non-verbal cues common to human conversation. Due to the absence of any real understanding, it is also difficult for the computer to effectively take the initiative. The second characteristic of human conversation is its frequently informal nature, which is understood within a shared context by partners that share similar levels of common knowledge about the world, conditions that are hard to achieve within human-computer conversations. Finally, a shared history is often a characteristic of human conversation that allows both parties in the conversation to interpret the meaning of certain utterances correctly, and without ambiguity.

Murray and Bevan (1984) argue that human behaviour is not simply goal-oriented, but has a social content that can be used to communicate the relative power and status of the speakers, or to influence the opinions of the other parties involved. In contrast,

speech-based interaction with machines can be characterised as being goal oriented, and limited to short phrases, with none of the richness of face-to-face human communication (Jaffe and Feldstein, 1970). These features also characterise human-human telephone-based conversation, which, lacks the richness of face-to-face communication, is devoid of non-verbal communication cues, consists of short phrases, and is often goal-oriented (Rutter, 1987). In this respect, human-human interaction by telephone shares many similarities with speech-based human-computer interaction. Human telephone-based interaction is therefore different from face-to-face interaction, and shares similarities with human-computer interaction, which further undermines any argument for modelling speech-based phone services on the principles of face-to-face human conversation.

Although guidelines exist for designing conversational interfaces, many of these are based on principles of human conversation, which cannot be applied to human-computer conversation due to limitations regarding the system's lack of real understanding, and inability to use non-verbal cues. Moreover, because telephone conversation is different from face-to-face conversation, it may not be appropriate to model automated phone services on face-to-face human conversation principles. Although Newell (1992) recommended that speech researchers should focus on dialogue design issues rather than the technological limitations of speech-recognition systems, technological issues remain important. In fact, recognition success is a key determinant of user acceptance and use of such systems (Baber, 1993). Technological issues will be discussed in the next section.

2.2.2.2 Limitations of speech-recognition technology

The difficulty of designing effective speech recognition systems was illustrated by Bristow (1986) by using a toothpaste analogy. In this analogy, he equates speech output with squeezing the tube to get the toothpaste out, whereas speech input is like trying to get the toothpaste back in again, which is exceptionally difficult. Despite intensive research few commercial speech recognition services have successfully deployed natural language capabilities (Gallwitz, Niemann and Noth, 1999), and where they have, recognition rates have been unacceptably low. For example, Jungk, Thull, Fehrle, Hoefl, and Rau (2000) evaluated a speech-recognition system designed to allow speech-based documentation of an anesthetic procedure in hospitals. Despite

the speech input vocabulary being limited and constrained to one, two, or three word commands, speech recognition errors were high. As a result of this, participants rated the system as being less controllable and complained about the length of the feedback dialogues. This study demonstrates that speech recognisers are still not capable of recognising natural speech in a fast, accurate manner, whilst also dealing with both interpersonal variations in input and environmental disturbances (Chan and Yeung, 1999). In addition, speech recognisers often fail in non-human ways (Rhyne and Wolf, 1993). For instance, due to the imprecise nature of speech recognition, and the absence of any contextual knowledge, there may be mismatches between what the user says, and what is recognised by the system. Such mismatches may be very different from the kinds of mistakes made by other people, and can be confusing for users.

Speech recognition systems are either speaker dependent or speaker independent. Speaker dependent systems require the user to train the system to recognize their voice in a process known as enrolment, which involves the user modelling the words that they will use, to provide the system with a template for their voice. Speaker independent systems do not have a template from a specific user to work from, and are designed to recognize a wide range of voices, including different accents and dialects. As would be expected, user-dependent systems have a far greater recognition accuracy (Fu, Chang, Xu and Pao, 2000). With respect to the actual spoken input provided by the user, single words spoken from a limited vocabulary are more successfully recognised than continuous input of multiple words (Alleva, Huang, Hwang and Jiang, 1998).

Current recognition systems are limited by the problem of recognizing input from unknown speakers, the difficulty in distinguishing multiple words as input, and the handling of large vocabularies (Noyes, 2001). This may explain why the majority of currently available speech-based automated telephone services are still based on an enumerated menu style, which constrains vocabulary leading to high recognition rates, but lengthens interaction times, and offers a less human style of interaction. There are, however, a number of usability problems associated with these enumerated automated phone services, and these will be discussed in the following section.

2.2.3 Usability problems of automated phone services

Automated telephone services have long been criticized for their poor usability (Yankelovich, Levow, and Marx, 1995) for two main reasons. Firstly, they are usually structured hierarchically, but contain little navigation information, often leading the user to become lost in the menu structure (Rosson, 1985; Wolf et al., 1995). Yankelovich et al. (1995) reinforce this view by commenting that ‘These [telephone-based] interfaces, however, are often characterised by a labyrinth of invisible and tedious hierarchies which result when menu options outnumber telephone keys or when choices overload users’ short-term memory’ (Yankelovich et al., 1995, p. 369). This problem arises from a fundamental difference between these speech interfaces and graphical user interfaces (GUI). A GUI can present multiple channels of information simultaneously using text, icons, menus, graphics, video and audio. These channels can be used as short-term memory aids and navigation cues, thus, providing the user with a set of options that can be scanned, and with permanent navigation markers if they get lost. For example, users can orient themselves by visually scanning graphics, drop down menus, clicking on linked pages, or looking at a site map. Such features are not available in speech interfaces, with speech being required to fulfil the dual tasks of information provider and navigation cue. Brewster (1997) believes that it is this dual requirement from a single channel of information that is at the root of the navigation problems.

Secondly, speech is produced in a slow and serial manner (Slowiaczek and Nusbaum, 1985). This means that the user must listen to the dialogue until the option they require is presented, remember the number corresponding to the menu option, and then interrupt the dialogue with a response if they wish. At the same time, users must also try to develop and maintain a mental model of the service structure and options so that, on subsequent uses, they will be able to interact more efficiently with the service by interrupting prompts earlier, and by using short-cut features such as type-ahead, so that they can enter the corresponding number sequence without waiting for the dialogues between their start and end points. However, forming a mental model of the service structure is difficult, as the numbers are arbitrarily assigned to menu options, and few navigation or structural cues are provided. The result of this serial presentation, and arbitrary assignment of number to options, is slow interaction and a burden on working memory (Schumacher et al., 1995, Sawhney and Schmandt 1998).

which can affect users to the point where they cannot successfully use the service (Kidd, 1985). This means that in practice many people experience problems and frustration using these services (Paap and Cooke, 1997). There have been three main approaches taken to resolve these issues: additional dialogue informing users of their position within the service; the use of earcons; and finally, the design of conversational interfaces. These approaches will be discussed in the following three sub-sections.

2.2.3.1 Additional dialogue

Rosson (1985) conducted a study to investigate the kinds of data structures and information retrieval methods that are most effective for remote information access using an automated telephone service. A hierarchically structured telephone service containing information about Austin, Texas, was simulated. The system used synthetic voice output, and required keypad input. Two different command interface styles were implemented to allow users to select items within a level, and also to move between levels. The first interface was a verbal mnemonic interface that provided commands that were matched to the lettering on the keypad keys, for example, the 'F' key moved the user **F**orwards through the hierarchy, and the 'B' key moved the user **B**ackwards through the hierarchy. The second interface was a spatial mnemonic interface where system functions were mapped onto a spatial arrangement of keys, thus utilising a spatial metaphor. For example, the '6' key was used to move right, and the '4' key was used to move left. Both systems allowed users to request to hear a repeat of the dialogue.

Participants were selected to represent a wide range of computing competency, from technical support staff to secretaries. All participants were asked to imagine that they were travelling to Austin, and that by calling the system they could access information about restaurants, shops, and entertainment venues. Performance data was logged using monitoring software, and attitude data was collected by using a post-task questionnaire.

The results showed a slight advantage for the verbal mnemonic interface for first time usage, suggesting that a verbal association between commands and keys was more useful than a spatial mapping. However, Rosson (1985) qualify this finding by

highlighting the fact that only one spatial arrangement was used (a cross shape), and that other different arrangements may be more useful. More dramatic findings were evident between individual users, with large differences between users in task retrieval times. On average, users made five separate wrong moves for each piece of information they tried to access. It was therefore clear that many users were applying an inappropriate navigation model, causing them to issue incorrect commands, leading to considerable difficulties navigating the hierarchical structure.

Information about the users' position within the hierarchy was implicitly provided by the service but required the user to understand the category-subcategory relationships between levels of the service. For example, after hearing 'Drinking' and moving to the next level of the hierarchy, the user would hear 'Bars', thus requiring them to infer from this relationship that they had moved one level down. Rosson (1985) identified this as a navigation problem, to which she responded by suggesting the provision of additional feedback dialogue to inform the user about where their previous menu selection had positioned them within the hierarchy, for example, 'You have moved from Drinking to the Bars category. The first bar is...' She acknowledged that this would slow down the interaction, and offered a number of alternative methods of addressing the problem, including integration of the navigation information into the dialogue, varying the syntax of the messages at each level, or by using a different voice to record the messages at each level.

Brewster (1997) argued that both the additional feedback, and the varied syntax, would lead to longer, more complex system messages, potentially obscuring the information being communicated, whilst the use of different voices may be hard to detect over the telephone network. This led to a trade off between the usability advantages to be gained from navigational cues, and the supposed reduction in usability caused by longer system messages embedded with navigation cues. Brewster (1997) proposed to address this situation by using 'earcons' as navigation cues for telephone-based interfaces, and this approach will be discussed in the next section.

2.2.3.2 Earcons

Earcons are abstract, musical tones that can be used in structured combinations to create sound messages to represent parts of an interface (Blattner, Sumikawa, and

Greenberg, 1989). They have previously been demonstrated to be an effective means of communicating information using sound (Brewster, Wright, and Edwards, 1993), and as powerful navigation cues in menu hierarchies for non-visual interfaces (Brewster, Raty, and Kortekangas, 1996). After designing a hierarchical system of earcons, they are then played in the background at each level of the hierarchy, enabling users to figure out their position both within a level and within the hierarchy. An additional advantage proposed is that the sounds would not interfere with the system dialogue, in the same way that one may simultaneously listen to the music and lyrics of a song. Earcons are also language independent sounds, which can therefore be used internationally.

Brewster (1997) showed that earcons could be successfully used as navigation cues in telephone-based interfaces if carefully designed to offset the reduction in sound quality caused by the use of telephone systems. Leplatre and Brewster (2000), used earcons to support navigation in menu-based interfaces where visual feedback was limited, for instance, on the small screens of mobile phones. Their results showed that menus enhanced by earcons led to performance advantages compared to visual-only versions of the interface. Thus, earcons appear to improve navigation of speech-based menus accessed using a telephone, and to enhance the navigation of visual menus on restricted displays especially on mobile devices. However, possibly due to the problems of background noise and degradation of sound over the mobile network, earcons were never successfully extended to the speech-based menu navigation of automated mobile phone services. The final approach to overcoming the usability problems of automated phone services has been the design of conversational interfaces, which removes the need to navigate a hierarchy of menu options, and to subsequently remember the options, thereby tackling both the navigation and the working memory problem associated with automated phone services. This approach will be discussed in the following section.

2.2.3.3 Conversational interfaces

One of the earliest attempts to design a more conversational style of dialogue was the PhoneSlave system (Schmandt and Arons, 1984). PhoneSlave was an automated telephone service allowing conversational interaction with an answering machine in order to retrieve specific voice messages. The interface was query based, and worked

by asking callers a series of questions, the responses to which were stored digitally, for example, 'Who is calling' and 'What is this in reference to?' Responses to these questions are then accessed by the system owner by asking questions in a prescribed sequence, such as 'Who left messages' and 'What did they want?' The PhoneSlave has no natural language understanding, it simply stores the messages in chronological order and plays them back to the owner, for example, when the user asks 'Who left messages' the system plays the sound file that was recorded when the question 'Who is calling' was asked. The system proved effective at eliciting voice message components from callers, its success being largely due to its ability to take the conversational initiative and to capitalise on user's tendency to follow conventional rules in response to questions. In this way user vocabulary was constrained and the system's limited intelligence was not exposed.

Several other research systems have used conversational paradigms bordering on natural language input to control live interactive systems over the telephone through speech recognition. SpeechActs (Yankelovich et al., 1995) was a telephone-based research prototype that integrated both speech recognition and speech synthesis, and was designed to allow travelling professionals to remotely access a number of applications including email and a calendar. The system was designed to be easy to use and easy to learn, with more direct access to information through the use of natural speech, rather than memorised commands. The system allowed the user to use phrases such as 'I'd like to check my mail please' or 'What do I have on my calendar the day after tomorrow.' By adhering to the conventions of human conversation, the interface was rated as highly usable. However, users encountered a number of problems with the speech recogniser, which was sensitive to variations in prosody and intonation. This led to a reduced speech recognition rate, and meant that dialogue flow was hard to maintain. A number of design recommendations were derived from the evaluations, including the provision of a barge-in function or a means of allowing users to speed up or slow down the speech, keeping the system dialogue brief and informative through the use of short cuts, and replacing some spoken prompts with auditory icons. Another important finding was that the terminology associated with the graphical versions of applications did not transfer well to the speech-based system, which highlighted a clear trend away from the technical use of language towards a more interpersonal informal conversational style.

MailCall (Marx and Schmandt, 1996) was a telephone-based messaging system, which aimed to improve on the efficiency of SpeechActs by organising messages into categories, and thereby simplifying the navigation process. As with SpeechActs the system was non-hierarchical with the entry point prompt asking the question ‘What else can I do for you?’ to which the user could provide a naturally formed request. When a recognition failure occurred at the entry point prompt, the user was asked to rephrase the request. If a recognition error occurred at any other point in the system, the user was prompted with progressively more explicit instructions in order to constrain the user’s vocabulary. Effective interaction was again hindered by high levels of speech recognition failure, with the designers suggesting that it may be necessary to constrain user responses by specifying that they must be short, simple requests.

More recent commercial speech recognition systems that can be accessed over the telephone line are the CSELT train timetable information system (Billi, Canavesio, and Rullent, 1998), the SpeechWorks air travel reservation system (Barnard, Halberstadt, Kotelly, and Phillips, 1999), and the Philips TABA train timetable information system (Souvignier, Kellner, Rueber, Schramm, and Seide, 2000). These systems tend to rely more on system-initiated or query-based dialogue in order to keep the speech recognition rate high. Telephone quality speech is significantly more difficult to recognise than a high quality stereo recording due to the limited bandwidth and distortions in the channel, and is made even more difficult when that speech is produced over a mobile phone network (Zue and Glass, 2000). According to Zue and Glass (2000) there are a number of issues that still stand in the way of natural language understanding conversational interfaces. These include the need to be able to detect and recognize new words, the need to expand recognition capability to a higher number of words, the need to be able to convey the capabilities of the system and the kind of input that it can cope with, and the need to isolate the source of error and offer effective recovery methods.

2.2.4 Implications for this research

This section has established that speech-based access to automated mobile phone services provides a convenient means of interaction, especially for mobile phone users

who are performing other visual tasks, such as walking around a city. Current guidelines for the design of automated phone services have been reviewed, some of which will be used to develop the mobile phone service prototypes tested in the experimental chapters of this thesis. A number-based style of interface was found to be the prominent style of interface for contemporary automated phone services. Such services cannot be classed as being conversational in the same sense as human conversation, but they are easy to learn, and lead to recognition rates that are high enough to be acceptable for convenient everyday usage. However, these services suffer from usability problems, which cause users to become lost in the menu hierarchies, and to forget the number associated with their desired menu option. Three techniques have been proposed to tackle these usability problems. Of these, earcons have not been successfully applied to menu hierarchies for mobile phone services, and the use of conversational interfaces has been hindered by poor recognition rates. The final technique, proposed by Rosson (1985) has not yet been formally evaluated. She suggested integrating navigation information into the service dialogue, varying the syntax of the messages at each level, and using a different voice to record the messages at each level. For the work reported in this thesis, all of these techniques will be used as a way of redesigning number-based mobile phone services through the use of interface metaphor, which has been a successful technique for improving the usability of graphical user interfaces. The following main section reviews the use of interface metaphor as a design technique in HCI, reveals that it has not previously been applied to the design of speech-activated mobile phone services, and builds a case for the potential usability benefits that interface metaphor may offer designers of speech-based mobile phone services.

2.3 The role of metaphor in user interface design

2.3.1 Introduction to metaphor

Benyon and Hook (1997) identified 3 ways in which navigation through an abstract information space could be supported: appropriate metaphors, using virtual reality and 3D interfaces, and using adaptive interfaces that accommodate individual differences in user's navigation ability. The approach taken in the work reported for this thesis was to use an appropriate spatial interface metaphor to represent the structure of the phone service information space in terms of a familiar structure from the real world.

The word metaphor is derived from the Greek meta (trans) + pherein (to carry) and means ‘transfer’, specifically the transfer of meaning (Ortony, 1975). A more recent definition is that provided by Lakoff and Johnson (1980) who describe metaphor as ‘understanding and experiencing one kind of thing in terms of another’. Metaphor is a mode of figurative speech, or a trope. Rather than providing a literal meaning, meaning must be deduced from the context in which the words are spoken, the meaning intended by the speaker, and from the listener’s knowledge of the world.

Philosophers as far back as Aristotle recognised the power of metaphor for explaining new ideas: ‘...for the receiving [of] information with ease is naturally pleasing to all: and nouns are significant of something: so that all nouns whatsoever which produce knowledge in the mind are most pleasing...but the metaphor in the highest order produces this effect. (Aristotle, 328BC/1995, p. 234)’

Rational theorists from the time of Plato (380BC/1955) through to Hobbes (1651) and in more recent history to Popper (1958) rejected the notion that metaphor had the ability to provide insights into one domain by borrowing concepts from another. They held that metaphor was merely a figure of speech, and a rhetorical device, marginalising its use to the arts and politics. Instead, the predominant belief was that only literal language should be used to explain scientific principles, because science and the truths about the world that it represents should not rely on context to be meaningful.

The first contemporary theorist to gain recognition for challenging the classical rational view of metaphor was Richards (1936), who introduced the idea that metaphor use is pervasive in language and that metaphor is a cognitive mechanism: ‘[metaphor is] the omnipresent principle of language [and that] we cannot get through three sentences of ordinary fluid discourse without it’ (Richards, 1936, p.92) and ‘Thought is metaphoric, and proceeds by comparison, and the metaphors of language derive therefrom’ (Richards, 1936, p. 96)

Richards (1936) also introduced a useful classification for the components of metaphor, and the relationship between them. He proposed the ‘tenor’ as the original

concept to be transformed, and the 'vehicle' as the second concept imported to metaphorically transform the 'tenor', although they are now known as the 'target domain' and the 'source domain' respectively. The 'ground' refers to the common features of the two concepts, whilst the 'tension' refers to the cognitive effort necessary to understand the dissimilarity between the two concepts. To illustrate this, a well-known metaphor, discussed in Lakoff and Johnson (1980), will be used: ARGUMENT IS WAR. In this metaphor, 'argument' is the target domain whose meaning is being understood by drawing metaphorical expressions from 'war', which is the source domain. The similarities between 'argument' and 'war' are called the 'ground', for example, an argument can be won or lost, the person one is arguing with can be seen as an opponent whose position is being attacking whilst our position is being defended, one can gain and lose ground, and adopt different attacking strategies. These shared features demonstrate that many of the things a person does when arguing are partially structured by the concept of war. However, there are dissimilarities, which can be referred to as the 'tension', for example, there is no physical battle when arguing, and achieving a win does not involve killing your opponent. Arguments and wars are different, one is verbal discourse, and the other is armed conflict. However, argument can be partially structured, understood, and talked about in terms of war, which, as long as the dissimilarities between the two concepts are understood, serves as an effective way of defining an argument.

This contemporary view of metaphor proposed by Richards was supported by an 'interaction' theory of metaphor proposed by Black (1962), and later adopted by other cognitive linguistic theorists (e.g. Kittay, 1987; Lakoff and Johnson, 1980; Schon, 1979; Reddy, 1979; Turner, 1987; MacCormac, 1976; Morgan, 1996). The first of these theorists to provide a detailed examination of the underlying processes of metaphor as the interaction of different concepts was Lakoff and Johnson (1980). They hold that it is not possible to understand, or to communicate an understanding of the physical world and the semantic world without using metaphor, and that 'Our ordinary conceptual system, in terms of which we both think and act, is fundamentally metaphorical in nature' (Lakoff and Johnson, 1980, p. 3). They essentially argue that metaphors are a conceptual construction central to the development of thought. Lakoff (1993) summarised the modern cognitive linguistic view of conceptual metaphor in a

number of ways important to the work reported for this thesis. Firstly, he states that metaphor is the primary mechanism through which we comprehend abstract concepts and perform abstract reasoning. Secondly, that metaphorical language may be used to explore the underlying conceptual metaphor from which the language derives. Finally, that metaphor allows us to understand an abstract or unstructured subject matter in terms of a more concrete subject matter. A detailed, and useful, definition of metaphor from this theoretical perspective, is offered by Indurkha (1992):

‘A metaphor is a description of an object or event, real or imagined, using concepts that cannot be applied to the object or event in a conventional way. The object or event being described is called the target, and the concepts that cannot be applied conventionally are called the source...the metaphor is made meaningful by interpreting the source unconventionally in the target. The unconventional interpretation can be arrived at on the basis of some underlying similarity between the source concepts and the target.’ (Indurkha, 1992, p. 17)

That metaphor is more than just a literary tool can be elucidated by examples of the way it has been used to both explain and extend scientific knowledge. Koestler (1964a) highlights the critical role of the ‘clock’ metaphor for research in cosmology, whilst Rose (1993) traces the various metaphors that have been used to explain the universe, including ‘the potters wheel’, ‘the chariot’, ‘the clock’, and ‘the engine’. A popular and well-known means of explaining the structure of an atom utilises the metaphor ‘an atom is a solar system’ (Gentner and Jeziorski, 1993). An increasingly common and relevant scientific metaphor for computers is that of ‘the brain’ (Hamilton, 2000a). Boyd (1979) uncovered many computer-based metaphors that have been used to explain mental processes in cognitive psychology, including ‘thought as information processing’, ‘information as encoded in memory stores’, and ‘remembering as an information retrieval procedure’. If such explanations were attempted in a literal way, they would prove to be difficult due to the abstract and complex nature of the concepts involved.

However, there is the danger that such metaphors may be over-extended, leading to inappropriate mappings and incorrect deductions (Nolder, 1991). A metaphor does not

draw exhaustive parallels between concepts, but instead provides a partial comparison between concepts, resulting in mismatches between the two concepts, and it is these mismatches that constitute the effectiveness of the metaphor by surprising and stimulating the listener (Hamilton, 2000b). These mismatches provide the basis by which metaphor can be exploited within HCI to help users to learn and effectively use a new system. The application of metaphor to computing interfaces is a technique known as interface metaphor, and the use of interface metaphors will be discussed in the following section.

2.3.2 Interface metaphors

The concept of metaphor as a guideline for designing human-computer interfaces gained popularity with the design of the Xerox Star graphical user interface (GUI) (Smith et al., 1982), which was the first incarnation of the desktop metaphor. A 'physical-office metaphor' was chosen to reflect the components of an office environment, which would have been familiar to the target users. The office desk was used as a cognitive and visual framework for the representation and organisation of files and folders on the computer. Users were encouraged to think about the interface objects, such as icons and folders, in physical terms. Since the early days of personal computing major software companies such as Apple have advocated the use of metaphor as a guiding principle in the design of interfaces (Apple, 1985, 1987, 1992): 'Use concrete metaphors and make them plain, so that users have a set of expectations to apply to computer environments' (Apple, 1987, p. 3) and '(User Interface) Metaphors are used to exploit specific prior knowledge that users have of other domains. For example the 'desktop' metaphor can be used to design an office information system. The use of metaphors can reduce the perceived complexity of user interfaces' (Apple, 1992, p. 19).

Within HCI, most psychological theories of metaphor consider it to be a learning aid (Mayer, 1975, Carroll & Thomas, 1982, van der Veer, 1990). The case made for the use of metaphors is that they reduce the time and effort necessary for new users to learn to use a system, by bridging the conceptual gap between the user and the system (Carroll & Mack, 1985). An effective interface metaphor is one that is appropriate, explicit, and quickly understood, and will lead the user to develop a mental model of the system that is closely related to the system image. This process involves 2 types of

mapping, the conceptual mapping of ideas, and the independent perceptual mapping of the representations of these ideas (Gaver, 1989). The continued need for metaphor arises from the increased complexity and functionality offered by computer systems, and the subsequent need to control this complexity to make the system easy to learn and easy to use. By designing the interface with reference to real world actions, objects, and procedures, users are being offered an increased degree of functional coherency, and more salient visual and auditory cues to help them to learn and understand the system.

Metaphors work by allowing prior knowledge from familiar domains to be applied to a new domain, which according to MacCormac is ‘the creative cognitive process of activating widely separated areas of long term memory and of combining normally unassociated concepts’, which leads to ‘an unusual juxtaposition of the familiar and the unfamiliar’ (1985, p. 147). It is therefore the interaction between conceptual similarities and dissimilarities that is at the heart of the metaphor process. The use of metaphor must involve some form of transformation; otherwise the construction is simply an analogy or juxtaposition and not a metaphor (Alty and Knott, 1997). With respect to computing systems, users must actively construct the relationships that comprise the metaphor during interaction with the system. The metaphor ‘seeds’ the constructive process through which existing knowledge is transformed and applied to the novel situation (Alty, Knott, Anderson, and Smyth, 2000). Hammond and Allinson (1987) argue that the metaphor should be as transparent as possible in order for the user to focus on materials and not the means of access to them.

Metaphor is viewed as an important technique facilitating the learning process, since it is easier to build up a completely new concept from other, more established concepts (Carroll and Mack, 1985; Smyth and Knott, 1994). This process is known as active learning, and Carroll and Mack (1985, p. 47) state that ‘metaphors can facilitate active learning...by providing abductive and adductive inferences through which learners construct procedural knowledge of the computer’. This means that the learner is encouraged to explore and make hypotheses about the problem space, which are either refuted or confirmed by subsequent interaction. Cornell Way (1991) supports this view, with Douglas and Moran (1983) coming to a similar conclusion as a result of their studies of typewriter learning. In their studies, typewriter users often made

predictions about how to use a word processor based on their typewriter knowledge. These predictions helped them to learn, but users were also frequently surprised by the differences in functionality between the two systems.

Active learning theory claims that dissimilarities between domains can facilitate learning by stimulating thought (Carroll and Mack, 1985). The active learning theory of metaphor draws a distinction between models and metaphors, whereby metaphors are incomplete and open-ended, and models are comprehensive and explicit. From an active learning perspective, accurate and literal models lead to passive learning because there are no deductions to be made (Jones, 1982), whereas metaphors require deduction to be made to facilitate the formation of mental models. Mental models and their relation to metaphor will be discussed in the next section.

2.3.3 Mental models

Norman (1988) provided a definition of mental models within the context of HCI: ‘the model people have of themselves, others, the environment, and the things with which they interact. People form mental models through experience, training and instruction’ (1988, p.17). Mental models of external events, situations, and systems help people to make predictions about outcomes before carrying out any actions. One of the key features of a mental model is that a conscious mental simulation can be ‘run’, and the predicted outputs can then be used to deduce conclusions about certain actions (Sheridan, Charney, Mendel, and Roseborough, 1986). According to Norman (1983), mental models are also characterised by a number of other features. They are invariably incomplete, as it is rare for a person to have a full structural and procedural understanding of a system. They are dynamic, and likely to evolve as a person has experiences that are relevant to the mental model domain. They can be internally consistent, but externally inconsistent, since a person may not have run simulations of the model to ensure its logical consistency, which can lead to many models being easily confusable. Finally, they are often based on intuition and superstition rather than on scientific fact.

Mental models are useful when performing three main processes (Carroll and Olson, 1988). Firstly, when learning to use a new system, a mental model may provide knowledge derived from analogous systems that the user has previously encountered.

Secondly, mental models can be used in problem solving, such as working out how to use previously unknown system functionality. Finally, mental models can be referred to when a user is attempting to understand unexpected system output, or when constructing a rationale for system behaviour.

The concept of schemata can be used to explain how mental models are constructed. A prominent theory of knowledge organisation is based on the concept of schemata, (Schank and Abelson, 1977), which are networks of general knowledge based on previous experience. Their function is to provide the knowledge necessary to allow people to behave appropriately when faced with everyday events and situations. When a person encounters an event, the relevant schemata are activated dynamically, and an appropriate mental model of the event is constructed.

Within cognitive psychology, Johnson-Laird (1983) provides an explanation of the structure and composition of mental models. Johnson-Laird (1983) argues that mental models consist of different types of knowledge representation in memory, either analogical, or a combination of analogical and propositional representations. Analogical representations are concrete and image-based, such as an image of a chair. Propositional representations are abstract and language-based, often making assertions, such as the sentence, 'the man is sitting on the chair'. Johnson-Laird (1983) makes an important distinction between the functionality of mental models and images. Whereas a mental model is usually constructed when a person needs to make a prediction about an event or situation, an image is considered to be a single representation. Preece, Rogers, Sharp, Benyon, Holland, and Carey (1994) invoke a useful analogy to clarify this by comparing an image to a single frame in a movie, whilst comparing a mental model to a short sequence from the movie.

There are three kinds of mental models (Carroll and Olsen, 1988): surrogates (Young, 1983), metaphors (Carroll and Thomas, 1982), and 'glass box' machines (DuBoulay, O'Shea, and Monk, 1981). A surrogate is a complete analogy of the target system that allows a user to formulate the correct actions in a given situation. According to Young (1983), it is very difficult for a user to construct such an analogy, even for a very simple system, which raises questions about whether people really do hold such mental models in their minds. A metaphor is a direct comparison between the target

system and a system that the user is already familiar with (the metaphor source domain). In contrast to surrogates, metaphor models are easy to construct, and provide explanations of a system's behaviour. However, one of the challenges of using metaphor models is that the designer must analyse what the user knows about the metaphor source domain (Young, 1983), as the system will be represented as the source domain. Glass-box models combine features of surrogates and metaphors, and have mainly been used as prescriptive mental structures. These models instruct users to think about the target system in terms of another system, whereas a metaphor invites the user to construct appropriate links between the two systems. It is metaphor models that are of interest to the work reported in this thesis.

Mental models may be either structural or functional, and Preece et al. (1994) describe a structural model of a system as being a description of 'how-it-works', and a functional model as a description of 'how-to-use-it'. A structural model describes the internal components and mechanics of a system, and is useful in allowing a user to make predictions about the behaviour of the system it represents. However, it does not account for how the user is going to perform a task with the system, and requires a large investment in effort to learn the model, and to use it. Miyake (1986) suggests that it is due to the difficulty in constructing such models that most peoples' understanding is based on functional models.

Functional models describe the way in which the system may be used, are based on knowledge derived from previous experience of a similar domain, and are context-dependent, which makes the model easier to use as it is tailored to a specific situation or event. When a system interface is metaphor-based, users will tend to develop functional mental models of the system. In such systems, the user develops a mental model of the system based on the source metaphor rather than on the way in which the underlying target system works. The metaphor therefore is the mental model that is learned. For the work reported in this thesis, metaphor-based functional mental models will be used, with the aim of providing the user with procedural knowledge of how to use a system by capitalising on their previous experience with a different domain.

To design a successful interactive system the designer must develop a conceptual model that is relatively complete and accurate, and that allows the user to form an accurate mental model of how the system works. Norman (1986, p. 46) states that ‘the problem is to design the system so that, first, it follows a consistent, coherent conceptualisation – a design model – and, second, so that the user can develop a mental model of that system – a user model – consistent with the design model’. The user develops a user model through using the system interface and its documentation, which are known as the system image. This model can be seen in Figure 2.1.

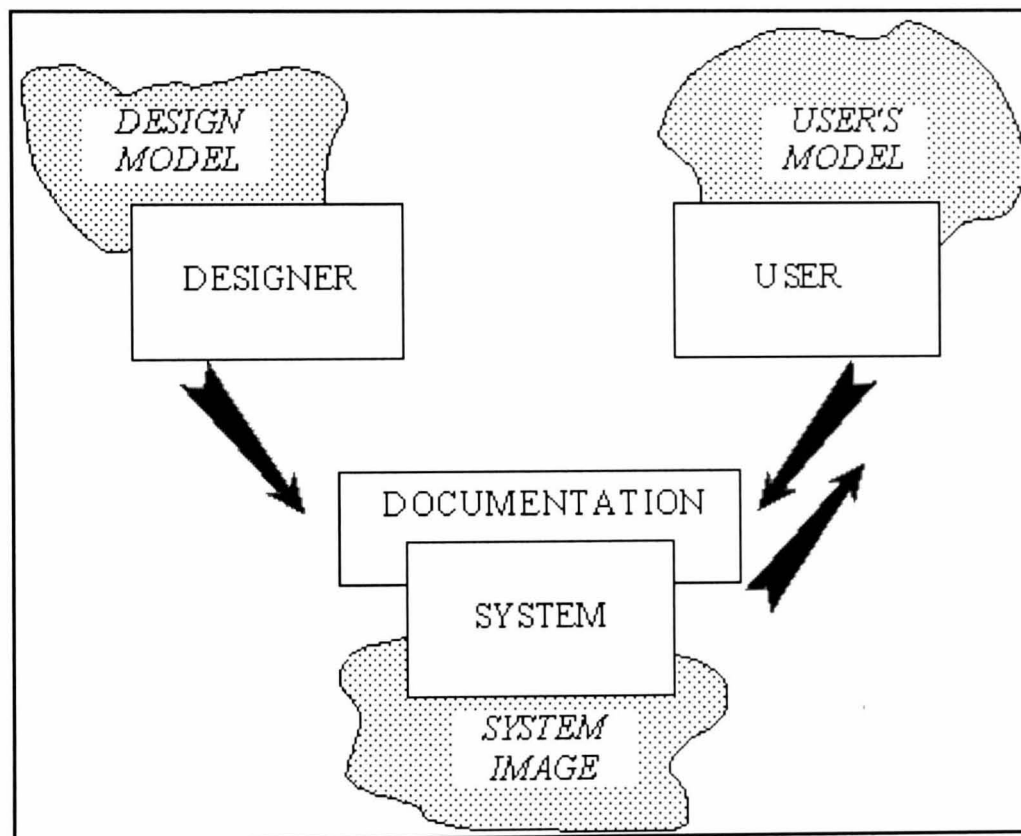


Figure 2.1. The design model, the user’s model, and the system image (Norman and Draper, 1986, p. 46)

The user’s mental model consists of internal representations of how the user interface functions. The system image is the interface that the user sees when they use a computer system, and is the implementation of the designer’s conceptual model. This must then be interpreted by the user’s mental model. With a metaphor-based system the designer has chosen to communicate their model of the system using a metaphor, which is often a simplified version of the original conceptual model (Carroll et al., 1988; Nielsen, 1990). The extent to which the system designer uses metaphor in the conceptual model, and consequently in the system image, will directly affect the

structure of the user's mental model. The goal for designers is to link two separate components of the user's mental model (Collins, 1995). The first is the component derived from interaction with the system, and the second is the component based on previous knowledge of interacting with or operating within the metaphor source domain. In so doing, the functionality and representation of the system overlaps with the functionality and appearance of the corresponding real world analog (metaphor source), enabling the user to develop a comprehensive and useful mental model of the system. For a speech-based automated phone service, the challenge is to convey a clear system image to the user, without using any visual representation of the interface.

For the work reported in this thesis, the system image will be represented as a metaphor-based verbal description, which requires the user to form an internal mental model of the system in the form of a visualisation. Visualisation is a cognitive activity (Ware, 2000), which leads to the formation of an internal mental model, or an image in the mind of the user. Visualisation can ameliorate cognition in three main ways. Firstly, by offloading some of the mental load from cognitive to perceptual processes, enabling cognitive inferences that would normally be done symbolically to be done using simpler and faster perceptual processes (Larkin and Simon, 1987). Secondly, through expanding the working memory available to a user by allowing some to-be-remembered items to be visually coded (Norman, 1993). Thirdly, by storing large amounts of information in an easily accessible form, similar to the way in which a map stores details of the environment (Card, Mackinlay, and Shneiderman, 1999). These benefits of visualisation may aid a user when interacting with mobile phone services by allowing them to dedicate more of their cognitive resources to the navigation of, and interaction with, their physical and social context of use. However, for visualisation to work effectively the mapping of the system information to the visual metaphor-based image must be appropriate (Card et al., 1999). A framework for incorporating metaphor into the design process, and for ensuring appropriate mappings between the system and the metaphor, will be discussed in the next section.

2.3.4 The metaphor design process

Carroll et al. (1988) identified three theories underpinning research on metaphor: operational approaches, structural approaches, and pragmatic approaches. Operational

approaches focus on the extent to which metaphors facilitate learning but fail to explain how this process happens in the mind. Structural approaches, such as the structure mapping approach (Gentner, 1983) interpret the metaphor process in terms of the mapping between the source and target domains, but see no benefits in mismatches. The final pragmatic approach recognises that mismatches between the domains will occur by definition, and that mismatches provide value by stimulating exploration. The pragmatic approach has been dominant in the understanding of the metaphor process, and active learning theory (Carroll and Mack, 1985) represents the most widely accepted view of the way in which metaphor functions.

Smyth, Anderson, Alty (1995) and Anderson, Smyth, Knott, Bergan, Bergan, and Alty (1994) proposed a pragmatic model of metaphor based on psycholinguistic literature, which is complementary to active learning theory. The model was developed from research on a series of prototype telecommunication systems as part of a European project called MITS (Metaphors for Integrated Telecommunications Services). This model draws on the concepts of system and user models (Norman, 1986; Fischer, 1991) and the terminology of system and vehicle suggested by Hammond and Allinson (1987) and utilised by Rogers, Leiser, and Carr (1988). They suggest that the activity at the interface can be described in terms of the intersection between two sets of functionality, that of the metaphor vehicle (M+) and that of the system (S+). Researchers involved in MITS later reflected on their experiences of designing and testing metaphor-based systems and incorporated the model into a six-step framework for engineering metaphor at the user interface. This framework will now be discussed, including a detailed explanation of the model, which comprises step three, and which forms its theoretical core. The six steps are:

1. Identify system functionality
2. Generate and describe potential metaphors
3. Analyse metaphor-system pairings
4. Implementation issues
5. Evaluation performed by observing users
6. Feedback on design

The first step involves identifying the functionality that must be provided to support work conducted in a specific task domain, leading to a functionality set 'S'. There are a number of task analysis techniques available within the HCI literature to achieve this, but the most well known is hierarchical task analysis (HTA). This technique is iterative, and involves breaking down tasks into sub-tasks and then into sub-subtasks (Shepherd, 1989). Information about tasks can be collected from a variety of sources including conversations with users, observations of users performing tasks, and operating manuals. The aim of HTA is to describe a task in terms of a hierarchy of operations and plans, the results from which are usually represented graphically.

The second step is to generate a number of metaphors that could support the functionality set 'S'. Approaches to achieving this include: monitoring the language used by prospective users for the use of metaphors; extending currently used metaphors; brainstorming by writing the system functionality on a board, selecting related items of functionality, and then mapping real world processes onto them; analysing the language used by customers when describing the kinds of systems they need; and finally, using ethnography as a means of generating descriptions of workplace objects and concepts that may suggest particular metaphors. The result of this step is a set of structured descriptions of each of the potential metaphors, called the Metaphor Set 'M'.

The third step involves examining the degree of match or mismatch between the functionality set 'S' and the metaphor set 'M', and enables the designer to assess the effectiveness of the mappings between a chosen metaphor and the system. There are four different categories of intersection between the system set 'S' and the metaphor set 'M' (see Figure 2.2), and these will be illustrated using the Macintosh wastebasket as an example.

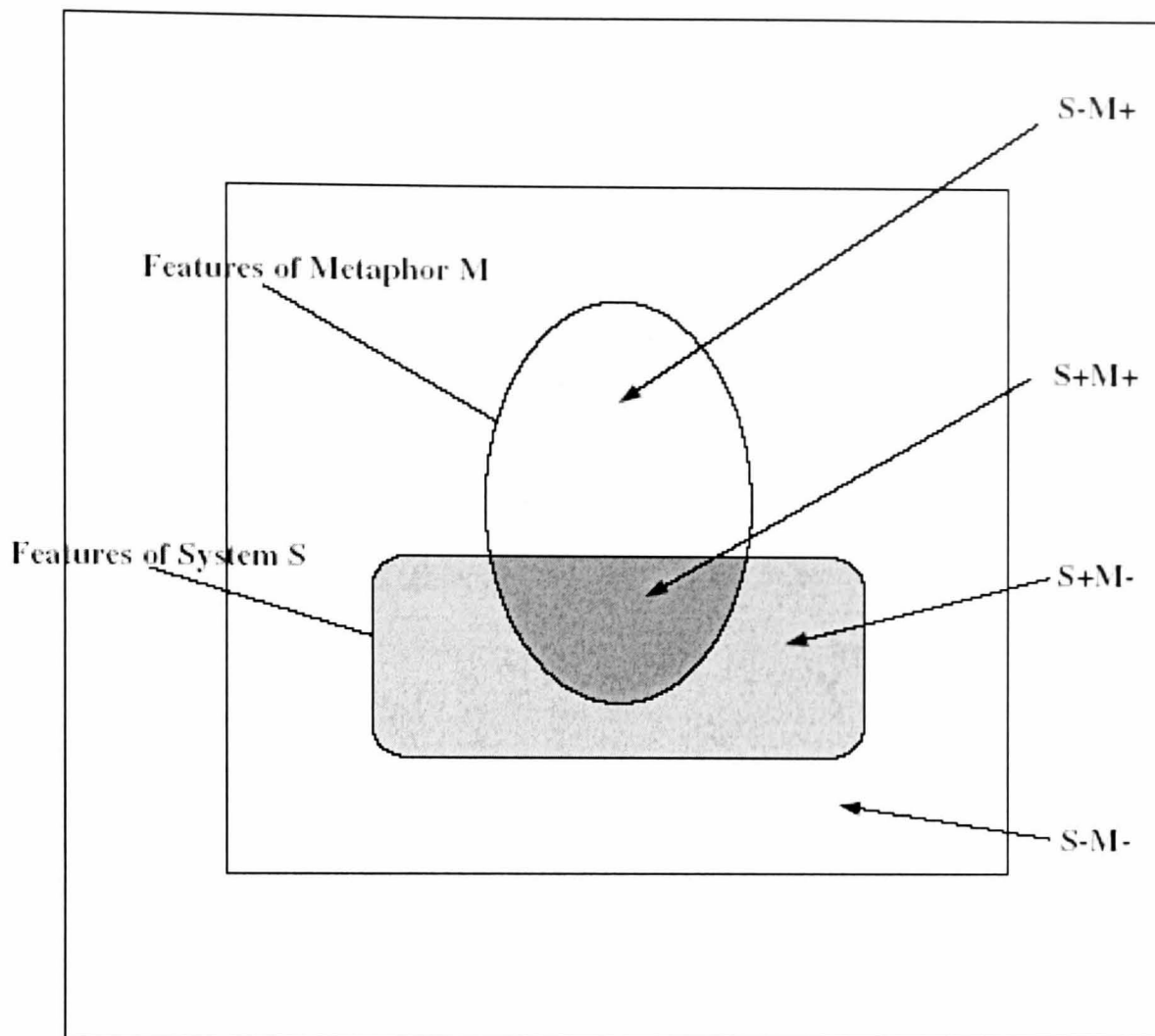


Figure 2.2. The interaction of the metaphor and the system in HCI

- S+M+ features are features of the system that map directly onto features of the metaphor. For the wastebasket, an example of such a feature is the mapping between ‘deletion’ and ‘throwing in the bin’;
- S+M- features are those features of the system that do not map onto the features of the metaphor. For the wastebasket, an example is the positioning of the wastebasket on top of the desktop, rather than underneath it, which is where it would be placed in the real world;
- S-M+ features are those features of the metaphor that do not intersect with the features of the system. These are features of the metaphor that cannot be applied to the system, and are referred to as conceptual baggage, leading the user to make incorrect inferences about the system. In order to avoid confusing users, designers should aim to limit these features in one of two ways. The first is to choose a metaphor with a narrow scope, and the second is to expand the system functionality to map more completely onto the metaphor. Of these, it would appear that the first reduces the need to rigidly adhere to the dictates of the chosen metaphor, leading to potentially unwanted functionality. An

example of conceptual baggage from the wastebasket is the way in which real world wastebaskets have a limited physical capacity, whereas the Macintosh wastebasket has a capacity limited by disc storage space;

- S-M- features are features that are common to neither the system nor the metaphor.

The fourth step concerns the issues of realism and consistency. It is essential that the metaphor is recognised immediately, but if the system too closely resembles the real world entity it can become an analogy, leading to incorrect expectations about system functionality. In terms of consistency, designers should be aware of the different interface styles that users may have gained experience with. By attempting to implement metaphors that are consistent with the look and feel of such interfaces, systems are more likely to be widely accepted by users.

The fifth step is evaluation of the metaphor-based system. Alty et al. (2000) recommend that evaluation take place in the user's normal working environment as a way of gathering ecologically valid data. In addition to observing and video recording users, they suggest gathering subjective data using questionnaires, and also performance data from user's interaction with the system. According to Alty et al. (2000), the key aspects to consider during evaluation are: cases of conceptual baggage, and the way that they affect user comprehension of the system; misunderstanding caused by S-M+ features, and irritations caused by S+M- features; inappropriateness of the metaphor; possibilities for the metaphor to provide new system functionality; and finally, the consistency and degree of realism of the metaphor. The final step of the framework is to ensure that the evaluation techniques and the data gathered could provide feedback that is useful for the redesign of prototype systems.

Alternative approaches to the process of metaphor design do exist, and the most comprehensive and structured of these is the five-stage methodology proposed by Neale and Carroll (1997). They recommend that the designer follow the following stages:

1. Identify system functionality

2. Generate possible metaphors
3. Identify metaphor-interface matches
4. Identify metaphor-interface mismatches
5. Manage metaphor-interface mismatches

A comparison of these five stages against the six steps of the Alty et al. (2000) framework reveals many similarities but few benefits. For example, steps one and two of the Alty et al. (2000) framework are similar to stages one and two of the Neale and Carroll (1997) methodology. In addition, step three of the Alty et al. (2000) framework provides a psycholinguistic theoretical-based approach to achieving the same results as stages three, four and five of the Neale and Carroll (1997) methodology. However, the Neale and Carroll (1997) methodology does offer useful advice as part of stage two ‘generate possible metaphors’. They suggest using visual techniques, such as sketching, to develop metaphors (Verplank, 1990). The designer produces a list of metaphorical words that describe an existing interface, and then sketches each of the words on paper. A list is then made of the desirable attributes and adjectives for the metaphor-based interface, which is further developed with reference to the sketches. The designer can then build on the words and sketches generated to sketch more extensive metaphors, which may later be developed into high fidelity prototypes. In order to make the process of generating metaphor-based words, attributes and adjectives more structured, the POPITS model proposed by Cates (2002) for the selection and implementation of graphical user interface metaphors may be useful.

Once a metaphor has been chosen, the key features of that metaphor need to be isolated and developed. Whereas Alty et al. (2000) proposed ‘fruitful conversation’ as a means of identifying additional components related to the underlying metaphor, and Neale and Carroll (1997) suggest thinking about related attributes and adjectives, the POPITS model (Cates, 2002) provides a more structured approach. Having identified the underlying metaphor, the POPITS model proposes systematically identifying the related **Properties, Operations, Phrases, Images, Types, and Sounds** arising from the metaphor, which are relevant to the system. Ideas generated within any of these six areas may also act as a catalyst to generate additional metaphors. It is claimed that this technique allows the designer insight into the user’s previous experience and

consequent expectations of the metaphor, thus enabling metaphor-based systems to be implemented that are more resonant with user experiences (Pugh, Hicks, and Davis, 1997). An example of the way the POPITS model may be used to generate features relating to a ‘Book’ metaphor is shown in Table 2.1.

Table 2.1. POPITS features for a Book metaphor

POPITS feature	Book-related POPITS features
Properties	Front and back cover, printed on paper, pages contain text, pages are numbered
Operations	Pick up a book, open a book, scan a page, reread difficult passages, turn pages forward and backwards
Phrases	Binding, pagination, read aloud, open, close, stop reading, underline, highlight, and speed-read
Images	Illustrations, smell of new paper, smeared pages, hard covers, and paperback
Types	Dictionary, atlas, science fiction, comic books, bible, Koran, law books
Sounds	Turning of pages, closing books, crinkling of a folded page, librarian’s stamp

The POPITS model appears to provide a useful means of developing the metaphors generated as part of a HCD process. However, because the model has only recently been formulated, there are currently no studies within HCI that have utilised it. In this respect, the work reported in this thesis represents the first use of this novel technique, especially in relation to metaphor design for speech systems. It is hoped that the model will allow relevant vocabulary to be generated that can subsequently be used for the metaphor-based descriptions provided by the service messages, and the metaphor-based options provided by the system prompts. In this way, the language of the service will be more meaningful and resonant with user’s experience of the corresponding real world metaphor source domain.

The Alty et al. (2000) framework offers a flexible theoretical-based approach to metaphor design and development that is also supported by practical experience in developing telecommunications systems, whereas the Neale and Carroll (1997) methodology has not been empirically evaluated. The Alty et al. (2000) framework was used for the work reported for this thesis, and was supplemented by the visual sketching techniques suggested by Neale and Carroll (1997), which were in turn supported by the POPITS model (Cates, 2002). Having established the overall

framework, and techniques, that will be used for the empirical work conducted for this thesis, the following section will review previous empirical studies of interface metaphor.

2.3.5 Empirical investigations of metaphor in computing

Since the development of Xerox Star's desktop metaphor (Smith et al., 1982), surprisingly little research has been conducted to provide empirical evidence for the supposed performance benefits of using metaphors for user interfaces. Mayer (1981) investigated the way that novices learn a programming language by providing some participants with a metaphor-based diagrammatic tool. The tool used a number of real world metaphors to explain specific concepts, for example, input as a ticket machine at a train station, and storage as a filing cabinet. Participants who were exposed to this tool before studying a training manual performed better at programming and recall tasks than those students who did not use the tool. Providing information about transactions using concrete real world metaphors improved performance, and this effect was most notable for low ability subjects, and when attempting creative programming tasks where the material was poorly organized. However, the study did not evaluate the effect of integrating metaphor into a computer-based interface.

Foss, Rosson and Smith (1982) investigated the effect of providing novice users with an advance organiser when learning to use a text editor. The organiser was based on the metaphor of an office containing filing cabinets in which folders and file were stored. The metaphor attempted to explain how the computer could be used to perform tasks such as creating, storing and retrieving files, for example, lines can be added or deleted from a computer file just as pages can be inserted and removed from a conventional file. In this experimental study, performance was measured as the number of tasks completed, the time taken to complete tasks, and the error rate. Results showed that exposing participants to an advance organizer that provided a metaphorical model for the operation of the system led to only slightly better performance. However, the authors suggested that the negligible effect of the advance metaphor-based organiser was probably a result of it providing insights into aspects of the system that were not directly tested by the tasks. This highlights the problem for designers of how to determine an appropriate metaphor-based model that will help users to learn the full functionality of a new system.

Amongst the earliest studies of metaphor were observational studies of users using the think aloud protocol (Mack, Lewis and Carroll, 1983; Douglas and Moran, 1983), and rather than analysing the effect of an explicit metaphor provided by designers, they analysed the effect of a metaphor that was voluntarily invoked by users. Think aloud protocol requires users to verbalise their thoughts and strategies as they use a system to complete a task, and has proved to be a useful tool for eliciting qualitative data about users problems and perceptions during human-computer interaction. These studies focussed on the way novice word processor users learned to use the system by analogy to a typewriter, and the findings were mixed. There were benefits from transferring knowledge gained from previous typewriter usage to learning to use a word processor. For example, participants had no problems typing text, nor were they confused by the fact that the characters they typed appeared on a monitor rather than on paper. However, users did find some features of the word processor surprising. For example, participants had problems figuring out how to set margins, as the mechanical controls that are provided on a typewriter to achieve this task were not present on the word processor. Overall, it was concluded that learners try to invoke their existing knowledge of typewriters to help them understand the semantics of word processor operations. Where there is a direct mapping between the functionality of the two systems, learning time is reduced, but for features that do not map, users often experience confusion and an increase in learning time. The study demonstrated the potential advantages offered by interface metaphor, but again failed to test the effects of an explicit metaphor-based interface against a non-metaphor interface.

Hammond and Allinson (1987) used a travel holiday metaphor to help users learn the navigation facilities of a computer-aided learning system. Within the overall travel metaphor, a number of sub-metaphors were integrated to reflect aspects of a real world travel system. The user could choose between 'go-it-alone' travel or a 'guided tour'. A map facility allowed the user to locate their position within the system, and finally an index provided keyword access to the system, in a similar way to the index of a guidebook. The performance of students was logged automatically, and students completed a questionnaire after usage session. The results indicated that the spatial metaphor helped students to learn and use the navigation facilities, and that the metaphor was also a useful aid for designers to think about additional facilities, such as the map facility that was developed for their study.

The desktop metaphor has not been empirically tested since its inception (Smith et al., 1982). However, graphical features of the interface, such as icons and menus, have been isolated and tested. In a study by Benbasat and Todd (1993) participants were assigned either to a direct manipulation interface condition in which icons were independently arranged on the desktop, or to a condition where icons were arranged into menus. Participants were required to perform tasks by formulating an overall goal, and then developing a step-by-step strategy to accomplish this goal. Task completion, and the time taken to perform the tasks, were measured. Results showed that, on average, participants using the direct manipulation interface performed tasks one second faster than those using the menu interface. The authors, therefore, argue that people using direct manipulation are more likely to engage in automatic processing than those using a menu interface, taking more motor time but less total time. This result supports the usability of the direct manipulation style of interaction provided by the desktop metaphor interface, but does not give any information about the relative benefits of a metaphor vs. non-metaphor interface. Potosnak (1988) reviewed studies comparing iconic interfaces with command line interfaces, and surprisingly found improved performance for the command line interfaces. However, such experiments are testing an interaction style rather than the use of metaphor as an interface design technique.

Smilowitz (1996) investigated the use of interface metaphor for a web browser. She compared two versions of the Mosaic World Wide Web browser, one of which used terminology based on a library metaphor, and the other of which was not metaphor-based. The terminology for the non-metaphor condition was simply based on Mosaic's function names, whereas the terminology for the metaphor condition was based on a library metaphor. For example, whereas the non-metaphor browser used the terminology 'open URL' and 'window history', the metaphor-based browser used 'search bookshelves', and 'stack of viewed books' respectively. User performance was measured as task completion, time taken to complete a task, and error rate. Subjective preferences were measured using a questionnaire. When using the metaphor-based interface, participants made fewer errors, performed tasks faster, and completed significantly more tasks successfully. In addition, participants perceived the metaphor-based interface to be significantly easier to use. Despite these results, she cautions designers that not all metaphors are good, and that whereas appropriate

metaphors can provide substantial usability benefits, poor metaphors may be no more effective than non-metaphor interfaces.

Kim (1999) evaluated the use of a spatial metaphor to help users with navigation aids provided within a cyber shopping mall. The experimental study investigated the navigation process of customers and gathered subjective evaluations about their online shopping experience. Two versions of a cyber shopping mall were implemented. One was based on a spatial metaphor, and the other based on a non-spatial metaphor. The results indicated that navigation aids based on the spatial metaphor were used more frequently, which led to a better understanding of the overall structure of the cyber shopping mall. In addition task completion and subjective evaluations were both improved in the spatial metaphor condition. This review of the empirical basis of the use of interface metaphor highlights a need for more empirical studies investigating and quantifying the specific advantages of using interface metaphor. It also shows that, of the empirical studies that have reported the benefits of metaphor, those evaluating spatial metaphor demonstrated the most conclusive results. The work reported in this thesis will attempt to contribute to the empirical evidence for the benefits of interface metaphor, specifically spatial interface metaphor. The following section will review the different categories of metaphor used for interface design, and present a justification for the use of spatial metaphor for the work reported in this thesis.

2.3.6 Categories of metaphor

Hutchins (1989) proposed three types of metaphor describing different aspects of the human-computer interface. Activity metaphors refer to the user's high-level goals, and structure expectations relating to the outcome of the activity, for example, writing a paper, or playing a game. Mode of interaction metaphors concern the user's perception of the computer, for example, whether they view it as a conversation partner, or an archiving tool. Task domain metaphors provide a structure to help the user understand computer-based objects and operations, for example, the way in which a file may be created or deleted.

Condon and Keuneke (1994) revised the classification proposed by Hutchins (1989), basing it on the underlying metaphor, rather than the medium in which it is presented.

This distinction is important because the presentation mode of a particular metaphor can change. An example of this is the way in which a spatial metaphor can be presented as a verbal description as it is in some adventure games ‘You are entering the nightlife district’, the verbal mode thereby invoking a spatial mode of internal representation. Their classification is composed of three forms of metaphor: spatial, activity-based, and interactional. Spatial metaphors define 2D or 3D spaces in which interactions and activities take place, activity based metaphors define the actions that can be performed upon the information or people within the space, and interactional metaphors support specific forms of communication (Condon and Keuneke, 1994). Properties of each metaphor type may be present in any single interface metaphor.

Of these categories, spatial metaphors have been used most extensively in interface design, and are based on our everyday knowledge of Euclidean space (Dieberger, 1996), which includes place knowledge, route knowledge, and survey knowledge (Kim and Hirtle, 1995). Freksa (1991) even argues that spatial metaphors are better than other types of metaphor, stating that:

‘Our knowledge about physical space differs from all other knowledge in a very significant way: we can perceive space directly through various channels conveying distinct modalities. Unlike in the case of other perceivable domains, spatial knowledge obtained through one channel can be verified or refuted by perception through the other channels. As a consequence, we are disproportionately confident about what we know about space: we take it for real...the spatial domain can be used particularly well as the source domain for metaphors with a non-perceivable or abstract target domain’ (Freksa, 1991, p. 361).

The fact that spatial metaphors are used extensively, and are considered by some to be superior to other types of metaphor, may not be surprising, given that humans are skilled at accessing, navigating through, and operating within physical space and informational space (Benyon and Hook, 1997). There are similarities between the information space of a computing system and that of the real world. For the purposes of this comparison it is useful to consider an information space in terms of an information artifact, which can be defined as ‘any artifact whose purpose is to allow

information to be stored, retrieved, and possibly transformed' (Green and Benyon, 1996, p. 802). Information artifacts employ various symbols, from which users are able to derive information, and which therefore define an information space. An information space may therefore be computer-based, such as the Web or a spreadsheet, or real world-based such as an airport or shopping centre, suggesting that abilities derived from real world spaces may be transferred to the navigation of computing devices.

Spatial thinking is used to represent and manipulate information in learning and problem solving in many different domains (Pellegrino, Alderton, and Shute, 1983), and human memory is thought to rely on the spatial arrangement of items (Kuhn and Blumenthal, 1996). Some studies have even suggested that users like to organise computer-based information spatially (e.g. Barreau and Nardi, 1995, Malone, 1983; Kaptelinin, 1996). Malone (1983) conducted an interview-based study with professional and clerical participants in order to investigate where and why participants stored files and folders in the locations that they did. Results identified the important role that file location has as a reminding aid for users. Barreau and Nardi (1995) built on these findings by conducting two studies with managers in order to observe and understand how they organised, and retrieved information from, their electronic workspace. The managers were asked to provide a tour of their systems, and were videotaped whilst this was done in order to provide a visual record of their file and folder organization. Managers were asked a structured set of questions, which were transcribed for analysis. One of the key findings from the studies was that the users showed a preference for location-based search for finding files. In a location-based search a user selects a directory in which the required file is believed to be located, and then browses the list of files, or array of icons, until the file is located. This can be contrasted with a logical search in which a user enters keywords into the search utility of the application. The authors hypothesized that users preferred location-based filing because it was more cognitively engaging, and imparted a greater sense of control. Another key finding was that the positioning of a file was used as an important reminding function, with users in both studies positioning files that required urgent attention in prominent positions, such as in the centre of the desktop. A later study by Kaptelinin (1996) reported similar findings, which supports

the use of spatial orientation as a way representing information through the use of spatial interface metaphor.

Several different types of spatial metaphors have been identified, with Benking and Judge (1994) identifying six different types of spatial metaphor that have been used for interface design: geometric forms such as cubes and spheres; artificial forms such as cities and houses; natural forms such as landscapes and trees; systemic structures such as highway systems and pathways; dynamic systems such as molecular and galactic systems; and finally traditional symbol systems such as rock drawings and sand paintings. Of these, the artificial spatial metaphor of a city has proved to be particularly popular, and Dieberger (1994a) suggests that the reasons for this are related to our familiarity with navigating around cities using various navigation aids such as maps, verbal descriptions, makeshift sketches, signposts, and through asking other people. Benyon and Hook (1997) suggest that metaphors with less structure may be useful for some systems, as they stimulate users to impose their own structure, and encourage exploration. Examples of such metaphors are: a wilderness or a desert metaphor for establishing a clear navigation route; the night sky metaphor for mapping and clustering purposes; and an ocean metaphor for distinguishing surface information from deeply embedded information.

Spatial metaphors may therefore provide a means of integrating navigation aids and structural cues into computer-based information spaces, capitalising on users' well-developed spatial cognitive abilities, and their preferences for organising and retrieving electronic information spatially. Spatial metaphor appears to offer an effective technique to address some of the usability problems of automated phone services. The following section considers the types of metaphors that have been formulated for new computing paradigms, such as mobile computing, leading to a review of the study conducted by Dutton, Foster, and Jack (1999), which provides evidence for the effectiveness of spatial metaphor in supporting the navigation of telephone service menu hierarchies.

2.3.7 Interface metaphors for different computing paradigms

Some authors (e.g. Marcus, 2002) have argued that an interface based on the fundamental notion of files and folders, applications and data, all embedded in office

artefacts such as desks and manila folders is no longer appropriate to new technologies and interaction paradigms. New areas of computing such as Computer Supported Cooperative Work (CSCW) and virtual reality have extended the task environment from the individual's desk to entire organisations interacting together, often working in different locations. New metaphors for these areas of computing have been designed since the late 1980s, and include office objects and the contents of whole buildings. Completely new metaphors have also been formulated, based on real world referents that are considered to be more appropriate to the domain, for example, novel navigational metaphors for virtual reality, such as 'eyeball in hand', 'environment in hand', and 'flying vehicle control' (Ware and Osborne, 1990).

The advent of wireless, mobile and handheld technologies has led to new interaction paradigms such as mobile computing, ubiquitous computing, wearable computing, and affective computing. A paradigm is a particular way of thinking about the design of human-computer interaction, and these new paradigms are characterised by a move away from a WIMP (**W**indows, **I**cons, **M**ouse, and **P**ointer) interface towards other forms of interaction. Each of these paradigms has signalled a shift away from applications designed for the desktop, and towards those that: can be accessed whilst mobile (mobile computing); that have become integrated seamlessly into the physical world (ubiquitous computing); that are contained within the clothes that people wear (wearable computing); and that recognise and respond to the emotions of the user (affective computing).

The design of mobile systems, such as mobile phones, introduces scenarios of use that are far removed from the office environment. For the work reported in this thesis, it was therefore necessary to investigate whether office-based metaphors were still appropriate, or whether new source domains were required that could replace those bound to offices and desktops. For mobile devices, new metaphors have often been the result of the limited screen size available on the device, which limits the presentation of information. Laptop computers are often based on a book metaphor, for example the Apple PowerBook, whilst Personal Digital Assistants (PDAs), such as the PalmPilot, have been designed using a pen and notepad metaphor. Metaphors have even been formulated for the non-visual control of PDAs. Pirhonen and Brewster (2002) approached the problem of controlling a digital music player running

on a PDA through the use of gestural and audio metaphors. The aim was to provide a means of input and output that did not require a visual interface, thus allowing the user to focus their attention on the visual world around them, and not the device they are using. They utilised a method proposed by Harrison, Fishkin, Gujar, Mochon, and Want (1998) to map the natural gestures a user would make to the interface functions, relating physical directions to logical order, for example, sweeping the hand across the screen from left to right would start the next track, whereas sweeping the hand in the other direction would play the previous track. Users were provided with auditory feedback on their actions through the use of earcons (Blattner et al., 1989; Brewster et al., 1993). These sounds were spatialised, using stereo panning, on a horizontal line in front of the user to represent the display of a CD player, with the sound representing the next track appearing to the right, and the sound for previous track appearing on the left. Such gesture-based interaction represents a more convenient way for users to perform tasks in mobile environments. Another approach is through the use of auditory and speech-based systems.

Auditory interface metaphors have been successfully applied to the development of complex auditory environments for visually impaired users (Lumbreras and Rossi, 1995; Savidis and Stephanidis, 1995; Mynatt and Edwards, 1995) and to provide cues for the navigation of hyperlinked, hierarchical, and auditory versions of GUI interfaces (Mynatt and Edwards, 1992). These studies used auditory icons, defined as sounds designed to trigger associations with everyday objects, just as graphical icons resemble everyday objects (Gaver, 1989). These auditory icons had to be selected using a joystick and keypad, and were based on a conversational model of interaction whereby the hypertext structure of the system was mapped onto the conversation. The Lumbreras and Rossi (1995) study will be used to illustrate how this mechanism works. In their study, they used a conversational metaphor to provide access to information within a 3D hypertext-based auditory environment. Each information node of the hypertext was mapped to a certain speaker in a specific position in space, which could be activated by clicking on 3D auditory icons. Each link reflected several conversational characteristics, such as requests, acknowledgements or questions. When listening to a simulated conversation, if the user was interested in a specific topic, they could use their knowledge of the voice type and speaker position to point to that speaker. This gave the information content an anthropomorphic characteristic.

and without being aware of it, the user was in fact working within a hypermedia system, since the hypertext structure had been mapped to the conversation. Although useful for visually impaired users, such spatial metaphors rely on the user being able to clearly hear the auditory icon, and on having access to a joystick. These conditions are not, therefore, suitable for mobile information access, with its high levels of background noise, and requirement for hands-free information access.

Dutton et al. (1999) adopted a different approach towards the use of interface metaphors for speech-based systems. They used spatial interface metaphors for implementing hierarchically structured automated telephone services. In their experimental study, two experimenter-generated metaphors were used, a department store, and a magazine metaphor, to implement a telephone home shopping service to be accessed from a fixed line telephone using keypad input. The user was required to navigate from the top level of the service to the level of individual items, and then to select how many items they wanted to purchase. A standard service was used as a control, and was designed by assigning menu options to keypad keys, with the dialogue as literal and free from metaphorical associations to shopping as possible. The two metaphor-based services used metaphor-relevant figurative language and sound effects to encourage users to imagine that they were interacting within a virtual three-dimensional space.

For the department store version of the service, users navigated between floors for different categories of goods by using the elevator, and within floors for different items, which were located either to the left, to the right, or straight ahead when leaving the elevator. For the magazine version of the service, different categories of goods were organised on different pages, and individual items were featured as numbered pictures on each page. The user was required to turn the pages of the magazine, and to select a picture on a page, as if they were reading and selecting items from a real home shopping magazine. For both metaphor-based services the messages and menu options were locational, but the meaning of the mapping between keypad and options were different. An example from the second level of the service is the mapping of keys '1', '2', and '3' to options 'left', 'straight ahead', and 'right' for the department store service, whereas for the magazine service the same keys mapped to 'first picture', 'second picture', and 'third picture'. The sound cues used for the

metaphor-based services were intended to enhance the realism of the metaphorical 3-dimensional space, but were not themselves arranged in real 3-dimensional space. Examples of sound effects used included elevator sounds for the department store service, and the sound of turning pages for the magazine service.

The performance measures recorded were the successful task completion, correct navigation route, and error rate which was the incidence of users not responding to the system prompts within the timeout period. Subjective attitudes were measured using a usability questionnaire. Dutton et al. (1999) found significant improvements in task completion rates, and service navigation for the metaphor-based services, and a significantly more positive attitude towards the department store metaphor service compared to a standard menu service. The study demonstrated that the use of interface metaphor has the potential to produce an effective improvement in a speech output service. The work reported in this thesis sought to investigate whether these advantages of using metaphor would extend to an automated mobile phone service requiring speech input, which would provide a convenient alternative to keypad based input for mobile users. The following section considers the problems with metaphor, and the arguments against the use of interface metaphor.

2.3.8 Problems of metaphor

Due to the incomplete mapping of a metaphor (source domain) to its target domain, mismatches can occur between domains caused by the increased tension between source and target domains. Mismatches can violate users expectations, causing inappropriate inferences and subsequent emotive responses (Neale and Carroll, 1997). The most frequently cited example is that of the 'trashcan' metaphor used in the Macintosh interface. By dragging files and folders into the trashcan, the user is discarding them, although they may still retrieve them if the trashcan has not been emptied. However, the trashcan is also used to eject floppy discs, and by dragging the corresponding floppy disc icon into the trashcan, the disc is ejected. Therefore, for the Macintosh, the desktop metaphor is limited because there is no office equivalent of ejecting a floppy disc. This inconsistent use of the trashcan represents a mismatch, which causes confusion to users, who feel uncomfortable with the concept of dragging valuable work that they have saved into the trashcan, which they have previously used to delete files.

However, such problems generally occur when the metaphor is not explicit, causing semantic confusion to the user. Carroll and Thomas (1982) investigated the problems users experienced when learning to use a word processor. They found that most problems were caused by the users' expectations that the word processor would behave exactly like a typewriter, which of course it does not, otherwise it would be a typewriter. Mack et al., (1983) and Douglas and Moran (1983) discuss the problems resulting from mismatches between typewriter and word processor functionality. The space bar on a typewriter produces nothing on the paper, it simply moves the guide along the current line. However, on a word processor, hitting the space bar inserts a blank character on the line, leading an experienced typewriter typist to make incorrect predictions about the functionality of the space bar on a word processor. However, the designers did not consciously intend to use the typewriter as a metaphor, which resulted in the metaphor being implicit and unstructured, and leading to confusion and error for the users. If the typewriter had been intended as a metaphor, the salient similarities and dissimilarities would have been made more explicit. It is therefore important for the designer to decide which features of a source domain are to be considered salient and which are not if an effective metaphor is to be designed (Kuhn and Frank, 1991).

Many researchers have proposed that users often generate their own metaphors to help them learn a new system (Blumenthal, 1990; Carroll and Thomas, 1982; Carroll and Mack, 1985; Rumelhart and Norman, 1981; Smith et al. 1982). Therefore, 'designers of [computer] systems should anticipate and support likely metaphorical constructions to increase the ease of learning and using the systems' (Carroll and Thomas, 1982, p. 108). In this way 'a user can draw upon his knowledge about [a] familiar situation in order to reason about the workings of the...new system' (Halasz and Moran, 1982, p. 383). It appears that in order for the designer to maintain a degree of control over users expectations, explicit use of metaphor should be an important consideration in the metaphor design process.

Carroll et al. (1988) maintain that mismatches are inevitable features of metaphors, because if they did not exist, the source and target domains would be the same, and the comparison would be an analogy. They propose that the burden is on the user to solve the problem of the mismatch through inductive reasoning:

‘...elaboration of metaphors is the mechanism by which the problem solver capitalises on prior knowledge of the source domain. Mismatches...pose questions; they stimulate elaboration mechanisms to construct a new understanding, to accommodate and integrate knowledge gained from using the metaphor to the new object of knowledge...’(Carroll et al., 1988, p. 76)

Dieberger (1994b) also believes these mismatches to be a strength rather than a weakness, but adds the caveat that dissimilarities between domains must be well designed:

‘These mismatches of metaphors often are important factors of the force of the metaphor. Mismatches in the metaphor can help considerably making a system useful if mismatches are designed well. The user interface principle of forgiveness is particularly important in metaphor mismatches – it allows the user to explore those unfamiliar features of the system and by exploring them she easily learns to use them for her own benefit ’ (Dieberger, 1994b, p. 57).

One of the fiercest critics of metaphor is Nelson (1990), who identifies metaphors as one of three ‘elements of bad design’, and considers metaphorical interfaces as ‘using old half-ideas as crutches’ (p. 237). He objects to the literal use of interface metaphors designed to look and behave like their real world counterparts because this is contrary to the way in which a metaphor benefits the user. A real world metaphor should provide a non-literal comparison between itself and a computer system, rather than simply being a computer-based analog requiring no exploration. Hammond and Allinson (1987) and Johnson (1987) support this view, pointing to the potential of computer-based interfaces to transform real world activities, thus making them more efficient, rather than simply modelling real world systems and processes. Nelson (1990) also remarks that the similarity between the metaphor and the real world object it represents is too tenuous to be useful. However, by saying this he is recognising the main strength of metaphor, that the parallels between the source and target domains are incomplete, leading the user to explore and to actively learn about the system.

Nelson (1990) also argues that by adhering to the constraints of the real world, designers are forced to provide design solutions that violate design principles, citing the dual functionality of the trash can from the Macintosh as an example. Other researchers have cited the Macintosh 'trash can' to focus on the way that metaphors can force designers to introduce logical and cultural inconsistencies, arguing that in the real world it would be placed under the 'desktop'. However, such mismatches rarely seem to trouble users once they have experienced and rationalised them (Preece et al., 2002).

Another danger for metaphor-based design is attempting to use a source domain metaphor that has been badly designed. A good example of this is the virtual desktop calculator, which has been closely modelled on the physical pocket calculator. According to Mullet and Sano (1995), the physical calculator is badly designed, based on a poorly conceived conceptual model, with excessive use of modes and badly labelled functions. The resulting virtual calculator is harder to use than the badly designed physical calculator (Preece et al. 2002). A metaphor source domain must therefore be carefully selected and evaluated to ensure that the conceptual model is clear and appropriate for the system being designed.

Some have argued that metaphors are culturally biased. When considering the desktop metaphor, Chavan (1994) argues that most people in India do not own desks or folders and do not have much experience with them. They do have bookshelves however, so if Indian researchers had invented the equivalent of the Xerox Star, it may have been based on bookshelves, with books having chapters and pages. If Chinese researchers had invented modern computing, screens may now be displaying vertically unrolling scrolls rather than windows (Marcus, 2002). Metaphor is therefore culturally biased, and must be evaluated for cultural consistency when a system is being developed for distinct cultural groups.

Halasz and Moran (1982) claim that metaphor is a limited technique for building an abstract mental model, as all abstractions are ultimately based on linguistic metaphor. They also point to the dangers of users making inappropriate or invalid deductions about an interface metaphor, although this would suggest that the metaphor had been poorly designed. Blackwell (2001) argues that the success of graphical user interfaces

is a result of direct manipulation rather than the use of metaphor, and casts doubt upon the substantial performance claims made about metaphor. Indeed the desktop metaphor was the first graphical metaphor-based interface, and the first interface to use direct manipulation as its interaction style. Therefore, with both interface design technique and interaction style being confounded in its design, this lends support to his argument. This point is a valid one, and as such will be investigated further within the experimental work reported in this thesis by comparing metaphor-based interfaces, whilst keeping the interaction style constant.

Metaphor has been accused of stifling creativity and potentially limiting functionality by recommending that designers base new systems on well known technologies or familiar real world systems. An example is the book metaphor used by Gentner and Nielsen (1996) to put the Sun Microsystems documentation online. They reflected that the chosen metaphor had prevented them from considering useful functionality, such as a feature to allow the user to reorder chapters based on the relevance score attained from a 'search'. However, in order to avoid this problem, for the work reported in this thesis visual techniques were used as part of the design process, which may allow more creative metaphors to be generated that suggest additional functionality, rather than restricting it.

The final problem concerns a specific category of interface metaphor, that of spatial metaphor. For spatial metaphors there is also the problem of scalability when considering interface design (Dieberger and Frank, 1998). Scaling up spatial metaphors causes two problems. Firstly, in terms of visualisation it is hard to represent large spaces in small displays, as this either leads to miniaturisation of the display features, or introduces the requirement for panning and scrolling to be able to locate the region of interest. Secondly, it may be difficult to navigate to distant objects within the interface if relevant location information cannot be viewed on the current screen or display.

To alleviate this problem, various researchers have argued for the use of multiple metaphors (Rumelhart and Norman, 1981; Halasz and Moran, 1982; Hammond and Allinson, 1987; Staggers and Norcio, 1993; Williams, Hollan, and Stevens, 1983). It has been suggested that any interface under design will have more functionality than

can be provided by a single metaphor (Alty, 1993). Gentner and Gentner (1983) conducted an experiment requiring students to generate metaphors to understand electrical circuits, and found that no one metaphor was sufficient for understanding all properties of the circuit. Gaver (1995) supports this view by suggesting that metaphor mapping is always partial mapping, because a metaphor will never be able to support the full system functionality. Opposition to the use of multiple metaphors focuses on the possible breakdown of the internal structures of each metaphor, leading to user confusion, poor coherency, and misconceptions due to the reductive effect of the metaphors (Spiro, Feltovich, Coulson, Anderson, 1989). However, the question as to whether to use multiple metaphors will ultimately arise as a result of following a framework for metaphor design, such as that provided by Alty et al. (2000). The interfaces designed for the work conducted in this thesis were speech-based, and it may therefore be possible to avoid the problems relating to the scalability of spatial metaphors, which are problems derived from the constraints imposed by the display sizes of graphical interfaces.

These arguments presented against the use of metaphor, rather than deterring designers from using metaphor for interface design, should simply be used to highlight the potential dangers and limitations of metaphor use. However, some of the arguments have led researchers to consider alternatives to the use of metaphor, and these will be evaluated in the following section.

2.3.9 Alternatives to metaphor

In his critique of metaphor, Kay (1990) suggests interface agents as an alternative. In the real world, an agent is a person that works on your behalf to achieve some goal, for example, an estate agent will help you to buy and sell property, whilst a travel agent assists you to find a holiday destination. Computer-based agents perform a similar job, and a well-known example is 'Clippy' (a paper clip that has human-like qualities) which is part of the help facility of Microsoft's Windows operating system. Interface agents are generally recommended for novice users and have become a common feature of software, often represented as virtual shop assistants or butlers (Oren, Salomon, Kreitman and Don, 1990; Laurel 1990; Lieberman 1997). However, by implementing embodied agents that sometimes look like virtual humans, and

appear to have some human-like intelligence, agents are simply a type of metaphor based on the 'human'.

Nelson (1990) argues for 'the construction of well-thought-out unifying ideas, embodied in richer graphic expressions that are not chained to silly comparisons' (p. 237), and proposes this be achieved through virtuality design, which is the design of principles. These freely designed principles are not bounded to an overall image, as in metaphor-based design, but are plastic and redefinable. He states that there are only a few organising principles left to guide interactive software design, such as spreadsheets, databases, and windows. He proceeds to cite the first spreadsheet program, 'VisiCalc', as being based on a genuinely new principle. However, VisiCalc was based on the metaphor of the paper spreadsheet. Another of his principles, 'windows' is based on the metaphor of a physical window. This casts doubt on both the novelty and the usefulness of these principles, and highlights the fact that it is difficult to design a completely metaphor-free interface.

Nardi and Zamer (1993) concede that metaphors are a useful technique for some purposes but inadequate for the design of complex scientific, engineering and business application interfaces. They propose visual formalisms as an alternative, which they define as '...diagrammatic displays with well-defined semantics for expressing relations' (p. 5). They are based on simple visual objects that contain their own semantics, rather than the semantics of some other metaphorical domain, and include tables, graphs, plots, and maps. They argue that it may not be appropriate to devise an imaginative metaphor for the information-intensive display of systems where precise and clear expression of semantics is paramount, such as the design of an electrical power plant, a life support machine for the space station, and a system for fault diagnosis in orbiting satellites. They argue that the goal in such applications is to model the true application semantics rather than modelling them metaphorically. In addition, such systems will be systematically taught to potential operators, which means the benefits of metaphor for novice users do not apply to this class of applications.

Their argument does appear to be salient, and there has been theoretical support for the cognitive benefits of visual representations in formal reasoning tasks (Stenning

and Oblerlander 1995), and also empirical support showing positive effects on the mental models of programmers (Green and Navarro 1995). Moreover, visual formalisms may be suitable for a small range of work-based applications where usability goals, such as satisfaction and entertainment are not a usability issue. However, there are a number of problems with the approach. Visual formalisms emphasise the external representation of a graphical interface, but do not support the user in their formation of a mental model. The user will attempt to form a mental model of the application regardless of whether this has been supported by the designer (Carroll and Olsen, 1988). It may therefore be the case that, when operating systems based on visual formalisms, users are devising metaphor-based mental models to help with the interaction. Studies of user's mental models when operating such systems would need to be conducted to establish whether it was possible to achieve their aim of non-metaphorical interface design. In addition, there does not appear to be any contradiction between designing a system interface based on a clear expression of semantics using visual formalisms, and a focus on the mental model of the user. Finally, proponents of visual formalisms take issue with the way in which metaphor requires a system to be designed to 'be like something'. By doing this they are rejecting the way in which new systems can be learnt and remembered more effectively by association with what is already known. This review of the alternatives to metaphor seems to highlight the problem of designing metaphor-free alternatives, although the use of visual formalisms does offer potential for the design of some complex safety-critical systems.

2.3.10 Implications for this research

This section introduced the concept of metaphor, and explained how it has been adopted within HCI as an important technique for reducing system complexity and helping users to learn to use new systems. The underlying psychological processes of metaphor comprehension and use were explained in terms of active learning theory, and mental models. Frameworks and models for metaphor design were presented and then evaluated, and the framework used for the work reported in this thesis described. Previous empirical studies of interface metaphor were then reviewed, which revealed a shortage of empirical evidence for the proposed performance benefits of interface metaphor. Different categories of metaphor that have been used for interface design were then reviewed, with the spatial category being the most widely used, and

offering the potential to capitalise on human's well developed spatial abilities. An investigation of metaphors used for different computing paradigms led to a review of the study by Dutton et al. (1999) in which two different spatial metaphors were used to implement a keypad-activated automated phone service. Results showed performance improvements relative to a non-metaphor service, and this study provides evidence that spatial metaphors can successfully be applied to speech output systems. The Dutton et al. (1999) study was found to be the only experimental evaluation of interface metaphor for automated phone services. There have been no previously documented attempts to utilise a user-centred design process to apply interface metaphor to speech-activated mobile phone services, which underlines the novel nature of the work conducted for this thesis. Arguments against the use of metaphor were then covered, followed by a review of alternative interface design techniques, which revealed the difficulty in designing non-metaphor interfaces. The following section addresses the 'human' component of the model proposed by Eason (1991) by examining the individual differences of users that may have an impact on the speech-based mobile phones services developed as part of the research work conducted for this thesis.

2.4 Individual differences in Human-Computer Interaction

2.4.1 Introduction

The individual capabilities, preferences and skills of users are now widely recognised as crucial factors to consider when developing more usable systems. Individual differences are the differences in the resources that users bring with them to the task (Bryce, 2000), and can play a major role in determining whether people can use an interactive system to perform a task effectively (Egan, 1988). The range of human performance on computing tasks is much greater than on most other work tasks (Borgman, 1989). Controlled studies of user performance have found ranges of 7:1 for text editing (Gomez, Egan, Bowers, 1986), ranges of 10:1 for information retrieval tasks by novice users (Dumais and Wright, 1986), and a range of 50:1 for programming (Klerer, 1984). In comparison, for non-computing tasks such as the tasks a factory worker would perform, the range of performance is 2:1 for 95% of the population (Wechsler, 1952; Salvendy and Knight, 1982).

When using a computer system, the user must assess the system state and decide on a course of action purely based on the external display, which consists of abstract symbols (Norman, 1988). This approach contrasts with the physical interaction that allows users to take apart, and look inside, non-computing systems to understand how they work, for example, a car engine. Perception and interpretation of computer displays involves information processing, and as a result, individual differences in users' cognitive abilities and preferences become important in HCI (Benyon, 1993). Cognitive abilities are relatively stable human characteristics, which change very slowly over time (Carroll, 1993), and include the mental processes of perception, memory and information processing.

Dillon and Watson (1996) analysed over 100 years of research on individual differences, and examined the relationship between work conducted in cognitive psychology and current analyses of users in HCI. They concluded that there is a core set of basic cognitive abilities that can influence the performance of specific computer-based tasks in predictable ways, and that user and task analyses for systems design could be constrained and improved by making them cognitively compatible to specific user types. They argue that whilst common sense variables such as task experience and domain knowledge (e.g. Greene, Gomez, Devlin, 1986; Nielsen, 1993) will remain important for establishing context, psychological measures of individual differences, such as cognitive abilities, provide a far more rigorous and consistent basis for comparing users.

Borgman (1989) examined correlations between more than a dozen characteristics that contribute to individual differences in information retrieval tasks, such as technical aptitudes, personality characteristics, and learning styles, and concluded that system interfaces can be tailored to suit the characteristics of specific user groups. For example, findings suggested that people with high spatial skills tended to perform better in graphical interfaces where the interface objects were presented spatially. If an interface is built according to an explicit metaphor, it may therefore be concluded that the target users are crucial components of an interface metaphor (Carroll and Mack, 1985), and must be considered if a system is to be accessible to, and usable by different user groups.

Gillan, Fogas, Aberasturi, and Richards (1995) found that specific individual differences interacted with the ability to interpret metaphors. They found that users with high levels of computing experience identified a larger number of underlying metaphors in an interface, and also provided more abstract interpretations of the metaphors than those users with low computing experience. They also found that users with high spatial memory were better able to identify and interpret interface metaphors. Sein and Bostrom (1989) investigated user's abilities to learn to use a metaphor-based electronic mailing system, and found that users with high visualisation ability took less time and performed better overall than users with low visualisation ability. These studies highlight the need to incorporate knowledge about individual differences into the design of both metaphor-based and non-metaphor systems. The problem is that the majority of user interfaces are currently designed with only a generic, ideal user in mind (Chen, Czerwinski, and Macredie, 2000).

Possible factors that could predict differences in performance are, previous experience with a system or task domain, technical aptitudes such a spatial ability and reasoning, age, gender, and personality. Apart from personality, all have proved to be strong and consistent predictors of performance. Personality is concerned with a person's approach to the world in general, and can be classed as a strategy rather than an ability. Personality traits correspond to patterns of behaviour and modes of thinking that determine a person's adjustment to the environment (Atkinson, Atkinson, and Hilgard 1983), and these traits change little and slowly, if at all, over time. Morgan and Macleod (1990) investigated the role of personality in a study comparing user preferences for both direct manipulation and command line interfaces. The aim was to examine possible personality differences between participants who preferred to use a direct manipulation graphical user interface, and those who preferred to use the command line interface. In this study, two groups of participants were assigned to either a graphical user interface or a command line interface, and were required to complete a number of tasks. Personality was measured using the British Standardisation of the 16PF test (Catell, Eber, and Tatsuoka, 1970) and no significant differences were found between groups. The results showed no significant association between personality type and interface preference.

An earlier review of studies that attempted to predict performance from personality types (Koubek, LeBould and Salvendy 1985) found that the correlations between personality and performance were typically weak. Allen (1987) also reviewed a number of studies that had attempted to investigate associations between personality and performance, and concluded that there was no evidence that personality was a predictor of skill level in HCI. With respect to speech-based systems, the case for the inclusion of personality could be argued on the basis that users might project personality onto the system voice. Indeed, Reeves and Nass (1999) discovered a tendency for users to project personality onto a computer-generated voice, suggesting that users develop expectations of a voice derived from their experience with other similar personality types. In this way users' perceptions of the personality of a voice could interact with their own personality type to affect performance. For the work reported in this thesis, an examination of the association between users' personality traits and their attitudes towards the service voice was not a factor that was of specific interest to the investigation of the usability of spatial interface metaphor. In fact, the effect of voice type was controlled as a factor by using the same synthetic voice for all of the service prototypes. However, perceptions towards the service voice were recorded within the usability questionnaire as part of the measurement of service likeability.

For the work reported in this thesis, the individual differences that were measured at different stages of the experimentation were age, gender, previous mobile phone experience, previous fixed line telephone experience, previous computing experience, verbal ability, spatial ability, cognitive style, and working memory. In addition, as a way of assessing users attitudes towards social aspects of context of use, attitudes towards mobile phone usage in public places were also measured. The following sections provide a description of these factors and a review of previous studies in which they have been analysed, leading to an explanation of the reasons for their proposed relevance to the speech-based automated phones service evaluated as part of the work reported in this thesis.

2.4.1.1 Age

Age has emerged as being a strong predictor of user performance in a number of studies. Egan and Gomez (1985) conducted a text editing study to assess the effect of

age on performance, as measured by error rate and task completion time. They found that an increase in age was a successful predictor of an increased incidence of first try error rates, and of an increased execution time. After breaking the task into its constituent components, they found that age specifically affected older participants' ability to generate the correct sequence of symbols required to make the required editing changes. Elias, Elias, Robbins, and Gage (1987) investigated the effect of age on learning to use a word processing program. Participants were supplied with two forms of training material, an audiotape and a manual. They found that older participants took longer to complete the training package and were more likely to ask for additional help. In addition, when later examined about aspects of the program, older participants performance was poorer than that of the younger participants.

Studies conducted using different applications have also found age to be a predictor of performance. For example, Caplan and Schooler (1990) found that older adults performance was poorer when using a software drawing package, and in a study investigating the use of a Calendar and Notepad application, Zandri and Charness (1989) found that older participants required more training time and more help than the younger participants. Despite these studies showing conclusive effects for age on performance, there is the difficulty of age being confounded with other factors. For example, it may be the case that older users have more experience with the system being tested (Egan, 1988). Age may also have an impact on user's spatial memory, which may in turn be the factor affecting their performance (Kelley and Charness, 1995). In their text editing study, Egan and Gomez (1985) found a relationship between age and spatial memory. When the effect of spatial memory on performance was held constant, they found a reduced effect of age on performance. It is therefore important to be aware of the presence of confounding variables in any analysis of the effects of age.

Love, Foster, and Jack (1997) investigated the effect of age on user's performance with an automated telephone service. The telephone service was an automated on-line music catalogue with a hierarchically structured database. Menu items were selected using a telephone keypad. Participants were required to access the catalogue, to compile a personalized compact disc (CD) by navigating through the menu systems to specific tracks, and then to select the tracks for the CD. The two objective measures of

performance were the interrupt rate, and the silence rate, with a high interrupt rate indicating good performance and low silence rate indicating good performance. The results indicated that age contributed significantly to the variation observed in participants' performance, specifically, older participants interrupted the service prompts less frequently. These findings suggest that age may be a salient factor to measure as a potential predictor of performance with the speech-based phone service prototypes developed as part of the work conducted for this thesis.

2.4.1.2 Gender

Whitley (1997) conducted a meta-analysis of the literature on gender differences in computing, and found that gender differences in attitudes towards and performance with computer systems has been demonstrated in a number of studies. Gattiker and Hlavka (1992) found that it was possible to classify user attitudes towards computer systems as a function of gender, whilst Venkatesh and Morris (2000) found differences between males and females in their perceptions of the usefulness and ease of use of a new software system. User attitudes and performance were studied over a five-month period, and they found that the male performance patterns were more strongly influenced by their perceptions of a system's usefulness, whereas female performance was more strongly influenced by perceptions of ease of use.

Palmquist (2001) conducted a study examining the types of metaphors participants used to describe and explain the World Wide Web. A questionnaire was administered to gather information about participant's gender, and level of previous computing experience. An additional section asked the participant for their choice and description of a preferred Web metaphor. General families of Web metaphors, together with sample keywords or synonyms for those families, were provided in the questionnaire, and participants were asked to either select their favourite Web metaphor, or to generate their own. They were also asked to write a short description of the characteristics that they felt made the metaphor an appropriate choice. They found that gender affected metaphor choice, with females preferring the 'highway' metaphor significantly more than the other metaphor choices.

In the Dutton et al. (1999) study, described in section 2.3.7, the usability of different metaphor-based versions of a telephone homeshopping service was evaluated. The

effect of gender on both performance with and perceptions of the services were analysed. They found that women were significantly more positive than men towards the service implemented using a magazine metaphor. The authors suggest that this could be because women found the style of verbal presentation for the magazine service more congenial than the men, or that they may have had more experience and knowledge of ordering products from magazines. This result suggests that gender could be a contributing factor to metaphor preference, and that it may not be possible to make generalisations across gender about the effectiveness of specific interface metaphors. Gender may therefore be a factor influencing users' preferences for interface metaphors, and it is for this reason that gender was analyzed as a potential predictor of subjective usability as part of the work conducted for this thesis.

2.4.1.3 Prior telephone and computing experience

Within HCI, previous experience has emerged as an important predictor of performance with a system (Egan, 1988). Previous experience refers to the user's experience with the actual computer interface used to perform a specific task, for example, text-editing experience. Domain knowledge refers to the knowledge and skill related to a task domain, for example, a proficient typist may transfer their skills to a computer text editing task. Domain knowledge has generally been found to be a less important predictor of performance because such knowledge begins to benefit a user only after they have acquired experience of using a specific interface (Egan and Gomez, 1985), but for the work conducted for this thesis was important as all participants would be expected to have had some previous experience of using automated phone services.

Rosson (1983) found that previous experience with a text editor could be used to predict the number of lines edited per minute. In another study examining information retrieval strategies in a file-search environment, Elkerton and Williges (1984) found that previous experience was the strongest predictor of search times. This knowledge may either have been gained as a result of previous experience with the system in question, or through experience of a different system that has been transferred to the system in question. Chrysler (1978) conducted a study to investigate the effect of previous experience on the time taken to complete a range of programming jobs and found that experience was correlated with completion time. Egan (1988) states that

the effects of previous experience on performance are especially pronounced during the early stages of learning a new system, with small amounts of practice producing big performance gains.

Due to an absence of research investigating previous experience in relation to metaphor-based speech systems, it was necessary to consider research conducted on visual interfaces. One such study was conducted by Maglio and Matlock (1998) who investigated people's metaphorical conceptions of the World Wide Web (WWW). They analysed the language used by participants to describe their actions whilst performing a web-based task. This analysis was based on the concept of image schema, which are basic pre-conceptual structures that arise from our embodied experience. Overall, web users described their use of the web as if they had been moving from place to place within some internal landscape. However, novice users were more likely to mix such descriptions with language relating to the use of external controls, such as the keyboard or mouse. This suggests that previous web experience may impact user's metaphorical conceptions of a computer system.

Novice and experienced users of speech systems have different skills and interface preferences, and it may therefore be the case that domain knowledge acquired from previous mobile phone and fixed line telephones could affect perceptions towards and performance with the metaphor-based speech systems developed as part of the work conducted for this thesis. In addition, previous computing experience with different interactive computing applications, some of which are structured hierarchically, and may employ speech output, could provide users with relevant domain knowledge to transfer to their use of speech-based phone services.

2.4.1.4 Verbal ability

Verbal ability refers to the ability to comprehend spoken or written words, and can be measured using vocabulary and reading comprehension tests. Verbal comprehension is composed of three main sub-processes, which are the lexical, syntactic-semantic, and pragmatic processes. The lexical level of processing is an unconscious level at which the sounds generated by spoken words are matched against stored templates and concepts. The syntactic-semantic level is concerned with deriving meaning. Finally, at the pragmatic level of processing, interpretation of meaning takes place as

a result of a person's understanding of the context in which the utterance occurred. Verbal ability has not been found to be a consistent predictor of performance in HCI (Egan, 1988).

Greene et al. (1986) conducted an information search study that required users to formulate a single query in order to access information on a specific subject. Results found that verbal ability could be used to predict performance, but that these effects were not consistent across all experimental conditions. In one of the few studies to find any effect of verbal ability, Vicente, Hayes, and Williges (1987) trained participants to use a screen-based browser for searching hierarchically arranged text strings. Participants had to locate target texts from files within a hierarchical file structure that consisted of three levels, with a total of 15 files. Performance was measured as successful task completion, and as time taken to locate the target texts. They measured a range of user characteristics including verbal ability, and found that verbal ability could be used to predict performance on the experimental task of searching for an item from within the hierarchical file structure. However, a factor analysis revealed the main predictor of performance to be spatial visualisation ability, with verbal ability adding only slightly to the predictive power of the model. Other studies show less conclusive results, for example, Egan and Gomez (1985) found that verbal ability was not a predictor of text editing errors, but that it could be used to predict task completion times. However, they caution that this result was probably related to the nature of the task, which required participants to read an instruction manual.

With respect to automated phone services, the study conducted by Love et al. (1997), which was described in section 2.4.1.1, found that the performance of low verbal ability users was significantly poorer than that of high verbal ability users when performing tasks with a hierarchically structured music catalogue telephone service. That verbal ability is a stronger predictor of performance in speech-based systems than in visual interfaces is perhaps not surprising. In an automated phone service, system output is speech, which requires the user to verbally comprehend the information provided in order to use the system effectively. It is not possible for the user to rely on a visual source of information, which means comprehension of the spoken prompts, messages and control options is of paramount importance. It is for

these reasons that verbal ability was monitored as part of the work carried out for this thesis.

2.4.1.5 Spatial ability

Spatial ability can be defined as the ability to perceive spatial patterns or to maintain orientation with respect to objects in space (Ekstrom, French, Harman, and Dermen, 1976), and is a cognitive characteristic that allows users to conceptualise the spatial relationships between objects (Jennings, Benyon, and Murray, 1991). It is widely accepted that spatial ability is not a unitary construct, but rather that it consists of different sub-factors (Stumpf and Eliot, 1995). Reviews of factor analytic studies show a range of spatial factors (McGee, 1979; Lohman, Pellegrino, Alderton, and Regian, 1987; Carroll, 1993). However, Lohman's (1989) three-factor model, based on an analysis of data from 35 studies, is perhaps the most widely accepted. The three spatial factors are Spatial Relations, Spatial Orientation, and Visualization. It is the third factor, spatial visualisation ability (SVA) that has been found to be the most important predictor of performance in HCI.

Salthouse, Babcock, Skvronek, Mitchell, and Palmon, (1990) define SVA as the 'mental manipulation of spatial information to determine how a given spatial configuration would appear if portions of that configuration were to be rotated, folded, repositioned, or otherwise transformed (p. 128). For the purposes of the research work conducted for this thesis, the definition provided by Ekstrom et al. (1976) will be used: 'the ability to manipulate or transform the image of spatial patterns into other arrangements' (p. 173). One of the main features of visualization is that it requires a figure to be mentally restructured into components for manipulation, rather than manipulating the whole figure (Ekstrom et al., 1976). Spatial visualisation ability can be measured using the VZ-2 paper-folding test. On the basis of interviews with participants who had completed the VZ-2 test, Lohman (1989) observed that high visualisation ability people 'solve items on such tests by generating mental images that they can transform holistically' and that such individuals are particularly proficient at 'generating, retaining, and transforming mental representations...' (p. 346). Despite the widely accepted presence of sub-factors of spatial ability, many studies within HCI have simply used tests such as the AH4 test Part 2 (Heim, 1970), to take an overall measure of spatial ability. Recent experimental studies have even

suggested that clear distinctions between sub-factors of spatial ability cannot be made (Colom, Contreras, Botella, and Santacreu, 2001).

Spatial ability and spatial visualisation ability have been significantly cited as reliable predictors of good performance in HCI by a number of researchers (Vicente et al., 1987; Vicente and Williges, 1988). Vicente and Williges (1988) found spatial ability to be the best predictor of time taken to locate a target text in a retrieval system. They suggest that constructing spatial mental models is a crucial component of task performance when dealing with hierarchically structured information. Stanney and Salvendy (1995) came to a similar conclusion in a study in which they designed two user interfaces to compensate for the inability of low spatial ability individuals to construct visual mental models of the structure of a menu system. The interfaces successfully compensated for these individuals, leading to an improvement in information search performance. They suggest that a key task component causing the differences between high and low spatial individuals is the construction of a spatial mental model.

Spatial ability has also been recognized as an important factor in increasing the efficiency of interacting with hypertext systems. Chen and Rada (1996) conducted a meta-analysis of 22 experimental hypertext studies, and their framework for analysis consisted of three components: users, tasks, and systems. The meta-analysis of users focussed on a range of individual differences, including spatial ability. They discovered that high spatial ability users did not access site maps and tables of contents as frequently as low spatial users. They proposed that this result was due to the increased ability of high spatial users to form spatial mental models of the structure of the underlying information space.

In terms of visualisation ability, Vicente et al. (1987) conducted a study, which was described fully in the previous section, in which participants had to locate target texts from files within a hierarchical file structure. Spatial visualisation ability was one of a range of measures taken, and they found that spatial visualisation emerged as the strongest predictor of time taken to locate the correct texts, with low spatial ability users most likely to become lost within the file system. In fact, participants with low visualisation ability took twice as long to find information than those with high

visualisation ability. Campagnoni and Ehrlich (1989) also found that users with high visualization ability were better equipped to construct an internal mental model of hierarchical information architectures, and at using these models for orientation and for directing their navigation. An additional reason for the improvement in performance may be related to the way that visualisation ability determines an individual's ability to draw analogies and apply them to a new domain (Pellegrino, 1985), which is a necessary process when using metaphor-based systems. This ability may also help users to learn the structure of a subject domain more quickly. Other studies have found visualization ability to be an important predictor of learning and performance in several domains (Sein, Olfman, Bostrom, and Davis, 1993): vocational-technical training programs (McGee, 1979), solving physics problems (Larkin, 1983), extracting information from maps (Thorndyke and Stasz, 1980), and text editing (Gomez et al., 1986).

Automated telephone services are structured hierarchically, with information being presented in a serial manner, and it is therefore difficult to include spatial cues (Martin, Williges and Williges, 1990). It is for this reason that spatial ability has not previously been found to have any effect on performance with automated telephone services. For the work reported in this thesis, structural and navigation cues will be provided within the speech-based phone services through the use of spatial interface metaphors, which encourage users to visualise the information space to form a spatial mental model of the service. Spatial and visualisation ability will therefore be of critical importance when interacting with these services, because in the absence of an explicit visual interface representation, users must rely on their ability to internally visualize the system structure. It might be expected that a system requiring users to visualise the structure and features of the interface in order to navigate effectively would not be beneficial to all users, particularly low visualisation ability users who rely more on the propositional representation of information in memory. However, it could be argued that services based on spatial metaphors may also benefit such users through the additional verbal navigational cues provided by the spatial metaphor, which would be difficult to acquire from numbered menu structures regarding paths and locations within a service.

This review has demonstrated that both spatial and visualization ability have been shown to affect performance in a number of studies. Improved performance within a hierarchical system, which requires the user to navigate and keep track of their position within the hierarchical menu system, has been associated with the ability to form spatial mental models of systems, and difficulties in the formation of such models has been linked to system navigation issues. It may therefore be suggested that structural and navigation aids based on a spatial metaphor would be expected to enable more effective mental model formation, and more efficient navigation than those based on a non-spatial metaphor. The spatial metaphors used to implement the phone services as part of the work conducted for this thesis were designed to facilitate the formation of mental models of the service structure. It is for this reason that both spatial and visualisation ability were measured as part of the work conducted for this thesis.

2.4.1.6 Cognitive style

Cognitive style refers to an individual's preferred and habitual approach to organising, representing and processing information (Messick, 1976). The most widely studied and applied dimension of cognitive style is the field-dependent / field-independent dimension, and stems from work conducted by Witkin, Moore, Goodenough and Cox (1977). Field dependence describes the 'degree to which a learner's perception or comprehension of information is affected by the surrounding perceptual or contextual field' (Jonassen and Grabowski, 1993, p. 87). When approaching problem solving tasks, field dependent individuals take a passive approach, process information globally, and thrive in situations where analysis is provided for them (Witkin et al., 1977). These individuals are less likely to impose a meaningful organisation on a field that lacks structure, and have more problems learning conceptual material when relevant cues are not available. Field independent individuals tend to adopt an analytical approach, prefer situations that require them to structure their own learning, sample more cues inherent in the field, and use these cues more effectively to complete tasks.

If cognitive style affects the way users structure and process information, it may in turn affect the way a person learns to use a computer system. However, within HCI cognitive style has not emerged as an important predictor of performance in general.

but there is some support for its impact. Coventry (1989), investigated user strategies for learning to use the UNIX operating system, and found that, if participants did not know the correct command, field-dependent participants were more likely to ask for help without making any attempt at the task, whereas field-independent participants were more likely to attempt the task and make errors than ask for help. Fowler, Macaulay, and Siripoksup (1988) found evidence to suggest that specific styles of interface could be developed to match the cognitive style of an individual, and subsequently improve their performance. They found that field dependent participants completed tasks faster when the system provided an inflexible dialogue structure that was system-guided, and consisted of formal language content.

Much of the research into cognitive styles has been conducted within the field of hypermedia learning (Durfresne and Turcotte, 1997; Shih and Gamon, 1999). Rather than a linear path through the material, hypermedia offers the learner a multitude of possible routes by which to explore the subject matter. This nonsequential access allows learners the freedom to browse the information and structure and manage their navigation, but also relies on the individual's ability to exploit this freedom effectively. The non-linearity of hypermedia may not suit all users (Ford and Chen, 2000), and studies have been conducted to investigate the link between cognitive styles and learning outcomes in hypermedia. Results have been inconclusive, with some finding an association (e.g. Jonassen and Wang, 1993; Weller, Repman and Rooze, 1994) and other failing to find significant links (Liu and Reed, 1995; Wilkinson, Crerar and Falchikov, 1997).

Dillon and Watson (1996), in their historical overview of individual differences research, suggest a number of reasons for the often inconclusive or inconsistent finding from cognitive style studies, but note that these are not based on empirical evidence. Firstly, the dimensions that have been identified may be superficial and need revising to reflect true information processing. Secondly, that individuals may be capable of manifesting several different cognitive styles depending on the context in which they are using a system. Finally, that a specific style may be highly correlated with specific tasks.

Due to the linear way in which information is presented to the user within an automated phone service, it may be argued that cognitive style is less relevant to this domain, as users are restricted in the way they can navigate through and explore the service. This may be true, but cognitive style may be related to user preferences for the metaphors used to implement the phone services, specifically in relation to the wholist-analytic dimension of cognitive style. In her study of users' preferred metaphors for the WWW, described in section 2.4.1.2, Palmquist (2001) also investigated the effect of cognitive style. She found that field dependent participants used large, nearly equivalent, descriptions to illustrate the characteristics of their preferred Web metaphor with an emphasis on what the Web is. Field independents often described the Web metaphor in terms of what they could do with it. These differences found in the metaphorical models described by participants fit well with the differences suggested by Pask (1988). He suggested that field dependents 'holists', are more concerned with the whole picture, whilst field independents 'serialists', are more concerned with individual features. This study suggests that metaphors as models must be feature rich for the field independent individual, who prefer more concrete metaphors, ones that closely relate to the functions and features already believed to exist in reality. Metaphors must also be creatively global in their representativeness for the field dependent, who prefer large abstract elements of a problem, so a model that lets them build in the detail as their understanding grows would be preferable. The metaphor-based services developed for the work reported in this thesis encourage the construction of an overall mental model of the service structure, which may be more suited to the field dependent individuals. The wholist-analytic dimension of cognitive style was therefore measured.

2.4.1.7 Working memory

Working memory is concerned with the active processing and temporary storage of information and can be used for verbal reasoning and comprehension (Baddeley and Hitch, 1974). It is defined by Baddeley (1992) as:

'a multi-component model controlled by a limited-capacity attentional system, which we termed the Central Executive. This was supported by at least two active slave systems, the Articulatory or Phonological Loop that is responsible for maintaining and manipulating speech-based information and the Visuo-

Spatial scratchpad or Sketchpad, which holds and manipulates Visuo-Spatial information. (Baddeley, 1992, p.8).

Working memory therefore acts as a 'scratch-pad' for the temporary recall of information, and can be used to store information only for short periods of time. Working memory has limited capacity and the information it holds decays rapidly. There has been little study of the effects of working memory in HCI because the dominant graphical user interface style uses multiple channels of visual information to provide short-term memory aids for users. Although it is possible to become lost within a visual interface, there are many visual wayfinding cues available that can be scanned by the user for orientation purposes. In contrast, speech systems provide a single channel of serial information that cannot be scanned or browsed. The user must try to remember the service structure, the menu options, and their location within the service hierarchy relative to the entry point. This serial presentation of auditory information places heavy demands on working memory (Tun & Wingfield, 1997). It is for this reason that for the work reported in this thesis working memory span was measured to investigate its relationship with performance and attitude towards the services.

2.4.1.8 Attitude towards mobile phone usage in public

As well as being personal devices, mobile phones are also social artefacts, and their use is influenced by the social contexts in which they are used (Palen, Salzman and Youngs, 2001). A mobile phone may be used in a range of contexts, from trains to offices, and when used in such public settings people in proximity to the conversation are either voluntarily or involuntarily affected (Ling, 1996). Due to the relatively recent phenomenon of widespread mobile phone usage, social norms concerning appropriate and timely use of mobile phones have yet to be established, and individuals are left to decide for themselves what constitutes considerate usage.

In a study investigating the evolving perceptions and adaptation to social norms of new mobile phone users, Palen et al. (2001) draw on Goffman's (1959) theory of public personas or 'faces' to explain their results. They propose that a mobile phone conversation occurs concurrently in the physical space of the caller, and the virtual space of the conversation, each space having its own social codes of behaviour and

appropriateness. They suggest that each space necessitates the adoption of a different public face, and that the caller must decide whether the face that takes precedence is the one that is appropriate to their current physical space, or the face for the conversational space. This conflict between physical space and conversational space can be used to suggest a number of ways that the public use of mobile phones could be perceived negatively. Firstly, the mobile phone user has chosen to be behaviourally present in a space different to the physical space they are sharing with other people, which may be perceived as inconsiderate by those people. Secondly, the conflict between the social norms of the spaces can lead to the user sometimes violating the norms of the physical space in preference to those of the conversational space. For instance, if a user is sitting in a quiet office and receives a call from a friend who is enjoying themselves in a bar, they may choose to adopt an informal conversational style that is inappropriate to the formal office setting. Finally, the face that a user adopts when taking a mobile phone call is different to the one adopted previous to the call. This can lead 'bystanders' to perceive the users 'conversational face' as being false, and even to the conclusion that their 'physical face' was also false.

Another interesting finding from their study concerns the attitudes of the surrounding public towards the value of the overheard conversation. A common complaint was that the conversations were frivolous and inconsequential, which was largely a function of the listener having no understanding of the context within which the conversation was occurring, and only hearing half of the conversation. This finding is of direct relevance to the work reported in this thesis, because the services require the user to utter menu options that would make no coherent sense to people overhearing the interaction. Possible negative reactions from people in close proximity to users may therefore affect users' perceptions towards the services when they are used in public places.

Love and Perry (2004) conducted a study investigating the behaviour and views of bystanders in response to different mobile phone conversations, and found apparently disinterested bystanders were actually highly aware of the conversation and able to recall detailed aspects of it. They concluded that mobile phone use is closely related to a person's morality, and that people make moral judgements about the manner of mobile phone use that is socially acceptable. Although acknowledging that social

norms for mobile phone use may vary across cultures and settings, they propose a working set of social norms that a mobile phone caller may adhere to. The caller is expected to: assess the situation and adjust their call length, call volume, and call content accordingly; become as 'apart' from bystanders as possible; and finally, to appear slightly apologetic about their call, as if they were grateful to the bystander for tolerating it. The degree to which users follow such norms when interacting with the services developed as part of the work reported in this thesis, may affect their perceptions of the services. In fact, attempting to adhere to certain norms, such as reducing the call length, will also lead to performance effects, as the user will be forced to interrupt more service prompts.

One of the features of using a speech-operated system is the lack of privacy, allowing other people to overhear what is being said in the same way that a mobile phone conversation may be overheard. This contrasts with a visual display, which can largely be obscured by the user to afford higher levels of privacy. When using a mobile phone to access a speech-operated automated phone service, the user may have similar feelings and perceptions to those experienced when making a voice call. It is for this reason that perceptions towards other mobile phone users, and level of personal comfort when using a mobile phone in public were evaluated as part of the work reported in this thesis.

2.4.2 Implications for this research

The review of individual differences presented in this section has established that the individual cognitive abilities, preferences, and skills of users have been found to affect performance with and attitude towards systems within HCI. The majority of the studies reviewed tested users' task performance with graphical user interfaces, and it was therefore necessary to hypothesise about the potential effects of some of the individual differences on performance with a metaphor-based speech-activated mobile phone service. A rationale was presented for the measurement of the following individual differences as part of the work conducted for this thesis: age, gender, previous mobile phone experience, previous fixed line telephone experience, previous computing experience, verbal ability, spatial ability, cognitive style, working memory, and attitude towards mobile phone usage in public places. Of these individual differences it is expected that spatial ability, and visualisation ability, will have the

greatest impact on user's ability to formulate a metaphor-based spatial mental model of the phone service. The following section addresses the 'environment' component of the Eason (1991) model, explaining the importance of considering physical and social context of use as part of system evaluation.

2.5 Context of use

An interactive device or service operates within a context of use, definitions of which are both numerous and varied within HCI. Preece et al., (2002) define context of use as 'the circumstances in which the interactive product will be expected to operate' (p. 207), and they recommend that the following four aspects be considered: physical, social, organisational, and the technical environment. Schilit and Theimer (1994) define context as 'more than just the user's location, because other things of interest are also mobile and changing. Context includes lighting, noise level, network connectivity, communication costs, communication bandwidth and even the social situation, e.g., whether you are with your manager or with a co-worker' (p. 23). Dey, Abowd, and Salber (2001) define it as 'any information that can be used to characterise the situation of entities' and go on to clarify it as 'typically the location, identity and state of people, groups, and computational and physical objects.' (p. 106).

Dix, Rodden, Davies, Trevor, Friday, and Palfreyman (2000) offer a broad definition of context of use as '...including the network, the broader computational system, the application domain, and the physical environment' (p. 296). They analyse the concept further by approaching context of use as being the location of the user and device within some form of space, which may contain other devices and users with which the device may interact. This introduces a social aspect to the definition, as they go on to explain that using a device within a space may affect other users. Dourish (2004) proposes two interpretations, where on the one hand it is a technical term offering system designers new ways to conceptualise the relationship between human action and the computer systems that support it. On the other hand, it is also a notion drawn from social science, drawing attention to certain aspects of the social environment. It is clear from these definitions that both the physical and social aspects of an environment are important to the analysis of context of use.

Whiteside, Bennett, and Holtzblatt (1988) recognized the importance of analyzing context, and found that although many interactive systems exhibited high levels of usability during laboratory evaluations, this was not the case when transferred to the real world. However, there are considerable challenges involved with evaluating the usability of mobile systems, involving the many activities and demands that can occur simultaneously and randomly, which makes it difficult to model the interactions between the activities, domains of use, tasks and users (Maguire 2001b). The practical problems concerning observation, data collection, and the limited means of controlling variables, also exist (Kjeldskov and Skov 2003). Despite these challenges, support exists for the added value of field-based testing of mobile systems (Abowd and Mynatt 2000, Brewster 2002, Johnson 1998). These benefits have, however, recently been questioned (Kjeldskov, Skov, Als, and Høegh. 2004).

Kjeldskov et al. (2004) conducted a study that evaluated a context-aware mobile electronic patient record system prototype (EPR) in both a laboratory condition and a field-based condition. The system, used by nurses, was selected as being representative of a system that would normally be extremely challenging to evaluate in its everyday hospital setting, due to the highly mobile, intense and often-stressful nature of the work involved. To evaluate such a system in a laboratory would therefore be a much easier option. In the study, the laboratory condition was a modified usability laboratory, which was set up to resemble part of the physical space of a hospital ward, whilst the field condition took place in a hospital. Audio and video data was collected using a wireless camera attached to the handheld EPR system. Participants were all nurses, and the tasks they were required to perform were derived from an analysis of typical work routines. The nurses were encouraged to think aloud whilst performing the tasks. The results showed no advantage of the field-based evaluation in terms of the number of usability problems identified, and the authors complained that in the field-based evaluation they were not able to force the participants to use certain aspects of the functionality that were of interest. However, when participants in laboratory conditions are forced to use system features they would not normally use and to subsequently think aloud while using them, the evaluation may not reflect the true usage patterns of a system. More importantly, participants may simply be generating false positive problems. This may have led to the high number of problems identified in the laboratory condition in the Kjeldskov et

al. (2004) study. A further limitation of the study relates to its reliance on think aloud as a data gathering technique, which may have resulted in less language being generated in the realistic hospital condition, due to work pressures. Although the study does offer an interesting perspective, it certainly does not provide strong evidence against the usefulness of field-based evaluation that can be extended beyond the domain of healthcare.

Mobile devices are typically used in highly dynamic contexts, with other people occupying the users' physical surroundings (Beck, Christiansen, Kjeldskov, Kolbe, and Stage, 2003), and activities such as walking whilst operating the device causing interaction problems (Brewster 2002) and placing heavy demands on both working memory and the visuo-spatial resources (Garden, Cornoldi, and Logie, 2002). When a mobile phone is used in public, physical aspects of the environment that can impact on usability include visual distractions (e.g. real-world navigation) and auditory distractions (e.g. traffic noise). In addition, the presence of other people, whether their actions interrupt the user, and whether their presence affects the user's actual or perceived ability to perform the task may act as social constraints on the interaction. These social aspects of context of use were covered fully in section 2.4.1.8.

This section has demonstrated the value of conducting evaluations in real world situations, which enable the effect of context of use on the usability of a device to be captured. For the work reported in this thesis, mobile phone users will be accessing and interacting with speech-based phone services from a range of locations, and the context factors of interest are those associated with the physical location and social settings. Physical and social context of use were therefore evaluated as part of the work conducted for this thesis.

2.6 Conclusions from the literature review

This review has shown that, when using currently available enumerated speech-based mobile phone services, people tend to get lost in the hierarchical menu system, and forget which number is paired with the menu option they require, which may largely be attributed to their poor mental model of the service. These problems may be exacerbated when the services are used in busy mobile settings due to environmental distraction and pressures. This thesis evaluates the benefits of using spatial interface

metaphors to improve the usability of these services, which is an approach that has not previously been attempted. This approach will incorporate the techniques suggested by Rosson (1985) for addressing the navigation problems of automated phone services, and in so doing, will attempt to provide users with a spatial mental model of the service that can be visualised. In accordance with the ‘human’ and ‘environment’ components of HCI identified by Eason (1991), the effects of individual differences and context of use will form part of the evaluation of the effectiveness of the metaphor-based services. This thesis aims to improve the usability of speech-based mobile phone services, whilst keeping speech recognition rates high, and ensuring that such services are designed with careful consideration of the salient characteristics of both the users and the context of use that may affect interaction.

The three main research objectives of this thesis were:

1. To investigate whether different metaphor-based versions of a speech-based mobile city guide service can improve the usability of the service compared to a non-metaphor version of the service
2. To investigate the effect of private and public context of use on the usability of a metaphor-based version and a non-metaphor version of a speech-based mobile city guide service.
3. To examine the effect of individual differences on the usability of metaphor-based and non-metaphor versions of a speech-based mobile city guide service.

:: CHAPTER 3

Research Methodology

3.1 Introduction

This chapter describes the methodologies used in the conduct of this thesis. The chapter begins by presenting an experimental overview to explain the studies and experiments conducted, and the experimental designs and techniques used. The chapter is then divided into two parts on the basis of the two main methodologies used: a human-centred design (HCD) methodology used to develop and evaluate the prototype phone services; and an experimental methodology used to compare the usability of different versions of the prototype phone services. The planning and conducting of the experiments is then explained with reference to: the variables; the sample of participants; the design of the experiment; and the data analysis techniques used.

3.2 Overview of the programme of research

Two preliminary studies and three experiments were conducted. Two different methodologies were used to achieve the different experimental objectives: a human-centred design (HCD) methodology; and an experimental methodology. The two

preliminary studies formed the human-centred design process leading to the implementation of the prototype mobile phone services. The three subsequent studies were conducted using an experimental methodology, and were conducted (i) to compare the usability of different service prototypes (ii) to investigate the effect of context of use on their usability, and (iii) to evaluate the effect of users' individual differences on their usability, thereby addressing the three research objectives. The following sections will describe these methodologies, and the techniques and materials used for each.

3.2.1 Methodology one: Human-centred design

Usability is widely recognised as being a critical factor in the design of successful interactive products (Shackel, 1981; Eason, 1984; Whiteside, Bennett and Holtzblatt, 1988; Nielsen, 1994). In current usage, the usability of an interactive product refers to whether it is easy to learn, easy to use, and enjoyable from the user's perspective (Preece et al., 2002), and may be broken down into the following goals: effectiveness, efficiency, safety, utility, learnability, and memorability. A usable system is one that allows a user to learn how to use a system quickly, and to operate the system effectively with low rates of error, leading to improved user acceptance. A HCD methodology has been advocated as an effective approach to achieve system usability (Preece et al., 1994, Maguire, 2001a). HCD is concerned with making user issues central in the design process, carrying out early testing and evaluation with users, and designing iteratively (Preece et al., 1994). According to the ISO 13407 (1999) standard on HCD, there are five core processes that must be undertaken in order to develop usable systems:

1. Plan the human-centred design process
2. Understand and specify the context of use
3. Specify the user and organisational requirements
4. Produce designs and prototypes
5. Carry out user-based assessment

After the initial planning stage, the remaining four stages should be conducted iteratively until the usability objectives have been attained (Figure 3.1)

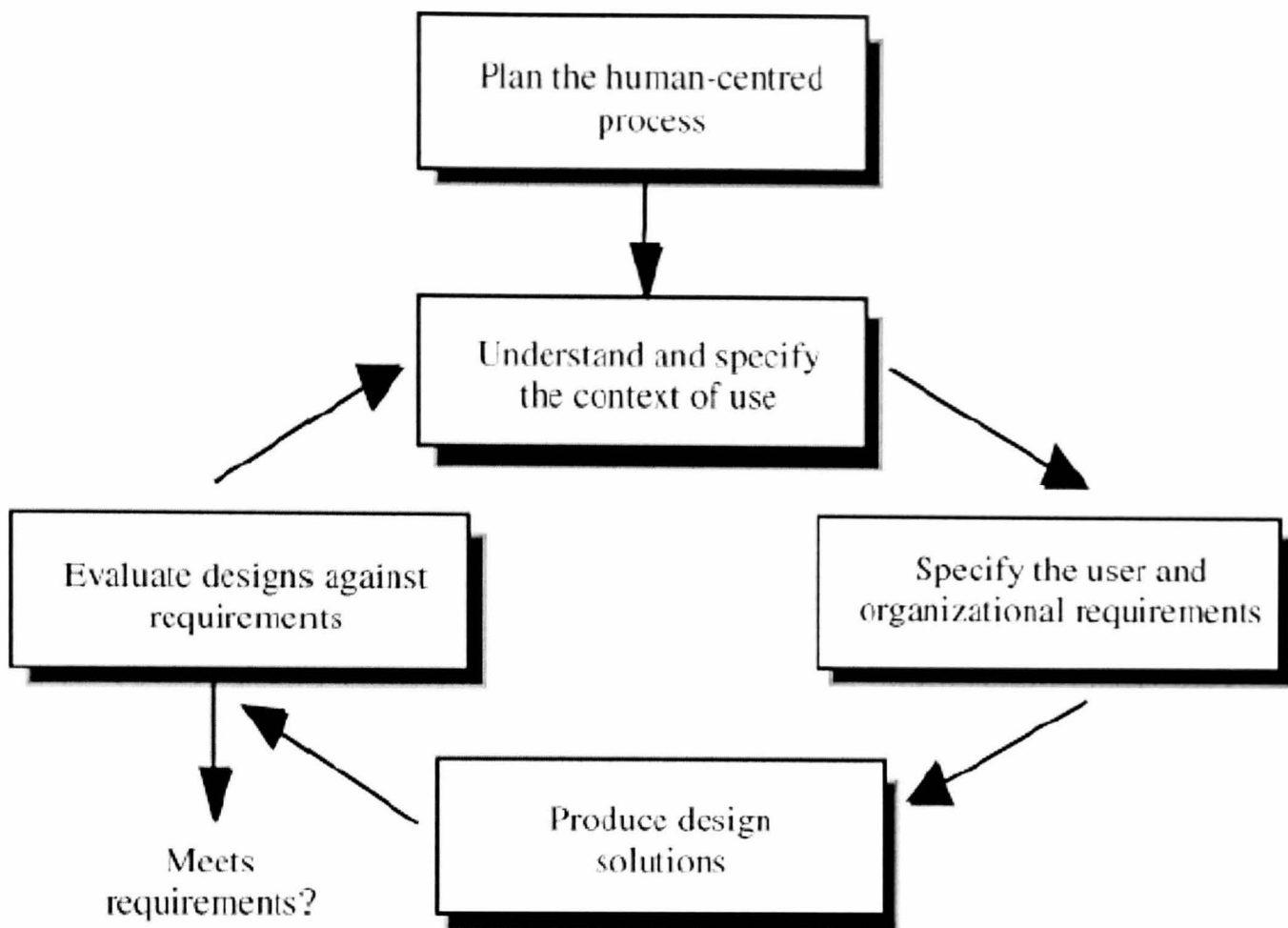


Figure 3.1. Human-centred design activities (from ISO 13407)

Table 3.1 was formulated by Maguire (2001a) and provides a summary of the methods and activities that can support each stage of the design process. The methods highlighted in italics are those chosen for the research conducted for this thesis. The sections that follow explain the importance of each stage of the HCD process, the methods used to design and evaluate the prototypes as part of the work conducted for this thesis, and the reasons for the selection of the methods used.

Table 3.1. Methods for human-centred design (adapted from Maguire, 2001a)

Planning	Context of use	Requirements	Design	Evaluation
<i>Usability planning and scoping</i>	<i>Identify and stakeholders</i>	<i>User requirements interview</i>	<i>Brainstorming</i>	<i>Controlled user testing</i>
Usability cost-benefit analysis	Context of use analysis	Existing system/competitor analysis	Design guidelines and standards	Satisfaction questionnaires
	Survey of existing users	User, usability and organisational requirements	Card sorting	Post-experience interviews
	Field study/user observation	Focus groups	Wizard-of-Oz prototyping	Participatory evaluation
	Diary keeping	Scenarios of use	Sketching	Assisted evaluation
	Task analysis	Personas	Paper prototyping	Critical incidents
		User cost-benefit analysis	Affinity diagramming	Heuristic or expert evaluation
		Task/function mapping	Software prototyping	
		Allocation of function	Parallel design	
		Stakeholder analysis	Organisational prototyping	

3.2.1.1 Stage one: Planning the human-centred design process

Maguire (2001a) suggests that at this stage the focus should be on gathering high-level information about the following:

- The reason for the system development, and the overall objectives
- The intended users, their capabilities and experience, and the tasks they will wish to perform
- The functionality needed to support the users
- How and why the system will be used
- The usability goals
- Guidelines that may be used
- User support
- Initial design concepts

These issues were covered as a result of the literature review and from discussion with other domain experts. The literature review helped to identify areas of weakness in current systems, and the potential usability benefits offered by interface metaphor for

future systems. In addition, the literature also provided descriptions of the functionality of current systems, including useful guidelines. Regular meetings between the principal researcher and three other HCI experts addressed issues involved with system functionality, potential users and their motivations, different scenarios of use, the usability goals, and the technical constraints.

3.2.1.2 Stage two: Understand and specify the context of use

Within the HCD process, context of use refers to the range of technical, physical, social and organisational conditions in which a system will be used (Maguire, 2001b). According to Maguire (2001b), for a relatively simple system, such as the one designed for the research conducted for this thesis, a context of use analysis can be informally conducted as part of discussions held at stage one of the HCD process. Context of use issues were therefore discussed during the planning stage of the design process. The data gathered from this analysis provides background information against which design and evaluation take place (Thomas and Bevan, 1995; Maguire, 2001b).

3.2.1.3 Stage three: Specify the user and organisational requirements

The system designed for the research conducted for this thesis was not designed to be used within an organisation, and therefore only the user requirements were specified. User requirements capture the characteristics of the intended user group, and the needs of those users in relation to the system being developed (Maguire 2001a). The result is a description of tasks the system needs to support, functionality that will support these tasks, potential scenarios of use, and possible interaction steps through the system.

In order to ensure that the system would support the type of work people would do with a city guide service, requirements were gathered from three sources: one-to-one interviews were conducted with five colleagues who all take at least two city break holidays per year; a review of city information in both the Michelin and the Lonely Planet guidebooks; and finally an interview with a consultant working for an online travel company specialising in European city break holidays. The three main themes that emerged from the interviews corresponded to the three main activities that were perceived to be most enjoyable when on a city break holiday, which were visiting

museums and exhibitions, eating out in restaurants, and nightlife activities such as drinking in bars, and listening to live music performances. These three activities were consequently matched to the menu categories at level two of the service, which were entitled 'Arts and Entertainment', 'Eating Out', and 'Nightlife'. Activities and interests within each of these three categories were then developed further to provide menu categories for levels three to five of the service.

Usability requirements were based on the ISO 9241-11 (1998), and were stated as being an improvement in effectiveness, efficiency and usability relative to currently existing number-based phone services. Effectiveness refers to how good the system is at doing what it is supposed to do, efficiency refers to the way a system supports users in carrying out their tasks, and the subjective usability of the system refers to the users' perceptions about how easy the system is to use.

It was also necessary to gain an understanding of the actions, or cognitive processes, a user is required to perform to achieve their goal when using currently available speech-based mobile phone services. This understanding is important when considering which features and functions are appropriate to new systems. In terms of user actions, current systems were analysed as part of the literature review conducted in chapter two (see section 2.2), to determine the system functionality that was available, and the subsequent user actions that were required to access this functionality. In terms of cognitive processes, the literature provided a review of the importance of cognitive factors such as working memory when using speech-based systems (Sawhney & Schmandt, 1998).

3.2.1.4 Stage four: Produce design solutions

For any system there are a range of alternative designs that can fulfil the system's specification, and it is the designer's job to explore this space of alternatives to identify the design that satisfies the system's constraints and goals as closely as possible (Preece et al., 1994). Users can be involved in testing design ideas by using experimental incomplete designs called prototypes, and this is an essential feature of iterative HCD, allowing designers to test their ideas and gather feedback from users.

In general there are two types of prototyping: paper-based and computer-based (Preece et al., 2002). Paper-based prototyping is low fidelity, includes techniques such as sketching and card sorting, and is a quick and inexpensive way of representing the system and of providing valuable insights early in the design process. Computer-based prototyping is high fidelity and provides a version of the system with limited functionality that the user can actually interact with. Prototyping can help designers to make design decisions by eliciting information from users about whether the system has the necessary functionality to support the tasks that the users need to perform. Information on operation sequences and user support can also be gathered, informing designers about the way users want to interact with the system, and whether the system supports them adequately during this interaction.

HCD should involve both low and high fidelity prototyping (Preece et al. 2002). For the research conducted for this thesis, the low fidelity phase involved rapid prototyping using card sorting, sketching, and paper prototyping of flow diagrams. These techniques were used within the two preliminary studies, and are described fully in chapter four (see section 4.2.1.2). The high fidelity phase involved evolutionary prototyping using Wizard of Oz (WOZ) prototyping (Fraser and Gilbert, 1991), which was used for preliminary study two, and all three main experiments. With rapid prototyping, the prototype is not developed into the final product, but provides information on the adequacy of different designs. With evolutionary prototyping, the initial prototype is constructed, evaluated, and after several iterations evolves into the final system design. During prototyping, guidelines were used to provide design guidance, and to ensure a degree of conformity with current systems. The specific guidelines used are discussed as part of the prototype design for preliminary study two, which is reported in chapter four (see section 4.3.2). As well as prototyping, two additional idea generation techniques were used within preliminary study one. Brainstorming was used to generate metaphors at the start of the design process, and the POPITS model (Cates 2002) was used to develop the metaphors generated. The POPITS model was discussed fully in section 2.3.4, and brainstorming is discussed within the methodology section for preliminary study one, in chapter four (see section 4.2.2). An overview of the different stages of design and evaluation, and the techniques used can be seen in Table 3.2. The following section describes WOZ prototyping, which was the primary prototyping technique used.

Table 3.2. The design and evaluation methods used for the experimental work

Study and experiment number	Metaphor design stage	Design solutions	Evaluation
Preliminary study 1	Generate metaphors	Brainstorming	Not conducted. The focus here was on the generation of ideas and new designs
	Select metaphors	Card sorting and sketching	
	Describe metaphors	The POPITS model	
Preliminary study 2	Implement metaphors	Paper prototyping using flow diagrams (+ Design guidelines)	Quick and dirty
		Wizard of Oz prototyping with 2/3 functionality (+ Design guidelines)	Quick and dirty
		Wizard of Oz prototyping with 2/3 functionality (+ Design guidelines)	User testing
Experiments 1, 2, and 3	Implement metaphors	Wizard of Oz prototyping with full functionality (+ Design guidelines)	Experimental method

3.2.1.4.1 Wizard of Oz prototyping

The WOZ technique involves a user interacting with a computer system whose functionality is actually operated by the designer or researcher (Maulsby, Greenberg, and Mander, 1993). The method takes its name from the book ‘The Wizard of Oz’ (Baum, 1900) about a girl who is swept away in a storm and finds herself in the Land of Oz. In this story the terrible Wizard of Oz turns out not to be a real person but simply a device operated by a man hiding behind a screen. The technique is less commonly referred to as the PNAMBIC technique and is an acronym from the scene in the film version of ‘The Wizard of Oz’ in which the true nature of the wizard is first discovered: ‘Pay No Attention to the Man BehInd the Curtain.’

To build a fully functional speech system is both time consuming, and requires high levels of expertise with the recogniser and synthesizer technologies needed (Klemmer, Sinha, Chen, Landay, Aboobaker, and Wanget, 2000). This makes it difficult to use an iterative design process, which requires the quick and repeated testing of prototypes that are inexpensive to produce, and do not require advanced technical competency. An alternative to building a fully functional system is to use the WOZ methodology (Fraser and Gilbert, 1991). WOZ involves the experimenter, or ‘wizard’, simulating that part of the system functionality that would be costly, time consuming, or require new technology to fully implement. While the user believes that he or she is interacting with a fully implemented system, the experimenter is allowed the extra understanding of the users that can be achieved through being a part of the interaction.

Fraser and Gilbert (1990) propose three pre-conditions that must be satisfied for WOZ simulations to be useful. Their first condition is that it must be possible for the human operator to realistically simulate the computer system. If a future application for a system requires the user to navigate through and manipulate objects from within a large database, but that database has not yet been implemented, it would not be possible for the user to simulate the large database of information effectively. However, if human capabilities are matched to the functionality of an application, then the application rather than just the interface can be simulated. For example, if the application required visual object discrimination which humans are skilled at performing, a simulation of a future version of the application could be successfully conducted.

The second pre-condition is that the designer should have a detailed knowledge of how the future system will behave in order to accurately simulate it. A danger here is that even when information about a system is known, such as speech recognition error rates, it can be difficult for the human operator to achieve the correct percentage error rate at the correct time within a restricted interaction span.

The final pre-condition is that the system should be convincing. To achieve this, the task demands should enable the operator to maintain the illusion that the user is interacting with a computer and not a human. This illusion is easier to maintain for text-based simulations, which provide only textual output that can be buffered a line at a time rather than at the wizard's typing speed. For speech output systems, Fraser and Gilbert (1991) mention the need to filter the wizard's voice so that it sounds mechanical, and for the human operator to strictly constrain responses to those that are matched to a system's capabilities. This latter point may be easy to achieve in prompted style dialogue systems, but much more difficult in systems that are designed to respond to a range of natural language responses.

Fraser and Gilbert (1991) provide a partial taxonomy of WOZ simulations into those that use natural language modalities such as typing and speech, and those which do not, such as the manipulation of objects in a graphical user interface, and keypad input. From the natural language simulations, a further sub-division can then be made

between simulations where only one of the interaction parties (wizard and participant) uses natural language, and where both interaction parties use natural language.

An example of a WOZ set up where only the participant used natural language was that performed by Hauptmann (1989) in which participants were required to manipulate screen images using spoken commands. On hearing the user's input, the wizard typed instructions to the computer in order to manipulate the screen images in the way requested by the participant. An example of a WOZ simulation in which only the wizard used natural language was a study conducted by Labrador and Dinesh (1984) in which participants were required to access a text-messaging service using a telephone keypad, and all system responses were produced as synthesized speech.

There are four categories of simulation where both the wizard and participant use natural language (1) the participant types and the wizard speaks (2) the wizard types and the participant speaks (3) the wizard and the participant types (4) the wizard and the participant speaks.

Examples of studies conducted in which the participant types and the wizard speaks are not reported in the literature although current interest in the design of systems allowing access to graphical user interfaces by visually impaired users may result in the design of such simulations.

WOZ simulations where the wizard types and the participant speaks are common, and involve mixing the modalities of speech and written communication. Most of these studies have involved simulations of listening typewriters with limited vocabularies (Gould, Conti and Hovanyecz, 1983; Newell, 1987), which require participants to dictate, whilst the wizard simulates the listening typewriter by typing words that then print to the screen.

WOZ simulations where both the participant and the wizard type are easier to set up and conduct, because in text based interaction there is no speech recognition problem. A direct correspondence exists between the key the subject presses and the character that appears on the wizard's screen. The wizard does not need to remember all of the legal words that may be recognized by the speech recognizer, and can instead just

pass the typed input through a filter to assess whether it belongs to the simulated systems vocabulary.

The final category of WOZ simulation involves both the participant and the wizard using speech, and is the style of simulation used for the work reported in this thesis, although the wizard activated voice prompts, rather than speaking the words naturally. Guyomard and Siroux (1988) conducted a study using a simulation of a phone-based Yellow Pages information service. As the automated service was new and was not replacing a human operated service, they did not have examples of human-human dialogue to work from. Therefore the first phase of their study was focused on requirements gathering, whilst the second phase was more constrained. For the first phase, user dialogue was either highly directed using prompts, or completely unrestricted using freely generated natural language. They found that infrequent users had problems with directed dialogue which required yes/no responses, whilst in the unrestricted dialogue condition the majority of user utterances contained hesitations and self-corrections.

The results from the second phase of the Guyomard and Siroux (1988) study uncovered two main problems, namely that dealing with user hesitations was very difficult, and that it was very difficult to predict the range of potential user input. These results suggest that the use of speech-based WOZ simulations may be more suited to a prompted style of dialogue with limited input vocabulary. In fact, McInnes, Jack, Carraro, and Foster, (1997) also recommend WOZ simulations for prompt and response systems which have a finite state dialogue and a fairly small set of input phrases.

In another study, Richards and Underwood (1984) conducted a spoken WOZ simulation of a telephone-based train timetable information service. The subjects used the service twice; being told on the first occasion that they were interacting with a computer, and on the second occasion that they were speaking to a human. On both occasions, the wizard provided the system output, but the voice was made to sound synthetic in the computer condition. A double-blind design was used in which the wizard was not told whether he was being presented to the participant as a human or a computer. The results revealed that the style and content of the participants'

utterances were affected by their perceptions of the system they believed they were interacting with. When told they were using a computer system, participants spoke more slowly, used a more restricted vocabulary, and formulated less ambiguous, more directed, phrases. These results suggest that participants should be led to believe that they are interacting with a future version of a computer system, so that their behaviour and utterances are representative of those that they would use when interacting with a computer system, and not a human.

Richards and Underwood (1984) also sought to investigate the effect of dialogue explicitness and politeness on the participants' responses through a range of different prompt styles. They found that when prompts were explicit and non-polite, participants produced the most concise responses, and that overall, politeness accounted for the largest proportion of redundancy in user responses. However, they warned that a reduction in politeness should not be at the expense of creating negative participant reactions to the system.

Fraser and Gilbert (1991) point to a number of variables that must be considered in relation to both the participant's and the wizard's performance with the WOZ simulation. With reference to the participant they must be able to recognize the wizard's responses, especially when generated as synthesized speech, which can be of very poor quality. For the work conducted for this thesis, synthetic speech was used over a mobile telephone network, which further degraded the quality of the synthetic speech. It was therefore necessary to conduct a number of pilot studies to assess the comprehensibility of different synthetic voices, leading to the voice used for the three main experiments reported. The participant must also be familiar with all of the vocabulary used by the system, which underlines the importance of using a HCD process to generate appropriate words and phrases, which was the process used for the research work conducted for this thesis. With reference to the wizard, the response time is an important variable. The wizard should aim to simulate the response time of the future system rather than the response time it would take a human. A related issue is that of training. In order to be able to respond quickly to participant commands, the wizard should undertake a training period to practice their role in the interaction. These issues were addressed as part of the work conducted for this thesis by running pilot studies to improve response time, and to practice the experimental procedure.

3.2.1.5 Stage Five: Evaluation

Evaluation is concerned with gathering data about the usability of a design or product by a specified group of users for a particular activity within a specified environment, or work context (Preece et al. 1994). A well-planned evaluation is driven by clear goals and appropriate questions (Basili et al., 1994). There are four main reasons for doing evaluations (Preece et al. 1994): to understand how users employ technology in the real world; to compare different designs; to test whether the product has reached the usability target; to ensure that the product conforms to a standard. Evaluation can occur at any point during the design process and should start as early as possible in order to check that the design supports user needs and aptitudes. Formative evaluations are conducted during the design process, whereas summative evaluations take place after the product has been developed to find out whether people can use the product successfully, and to make judgements about the finished product.

Any evaluation is guided by a set of beliefs, which may also be supported by theory, and these beliefs and the methods or techniques associated with them are known as an evaluation paradigm (Preece et al. 2002). There are four paradigms central to evaluation. The first is ‘quick and dirty’ and involves the designer talking to users about their design ideas to make sure they are in line with users’ needs and preferences. The emphasis here is on quick and informal feedback in the form of descriptions or anecdotes that can then be fed back into the design process. The second is ‘usability testing’ which involves measuring typical user’s performance when performing representative tasks with the system. Performance is usually measured in terms of the number of errors made, time taken to complete the task, and sometimes the route taken through a system. Users’ opinions are also gathered using questionnaires and interviews. The emphasis here is on experimenter-controlled conditions, typically in a usability laboratory, and on the gathering of quantitative data. The third paradigm is ‘field studies’, which are conducted in natural settings with the aim of understanding how a system impacts a user in their everyday context of use. Techniques such as interviews, observation, and ethnography are used to elicit qualitative data. The final paradigm is ‘predictive evaluation’, which involves experts using their knowledge of users to predict the kinds of problems users might encounter with a system. The evaluation is often performed against a set of cognitive principles.

For the work reported in this thesis, the two paradigms used were ‘quick and dirty’, and ‘usability testing’. The goals for the evaluation of the paper prototype flow diagrams, and the first version of the WOZ prototypes in preliminary study two were (1) to ensure that the designs supported the user requirements, and (2) to provide fast feedback to redesign the prototypes. The two paradigms that could fulfil these goals were ‘quick and dirty’ and ‘predictive’ evaluation’. Rather than using the more structured approach of predictive evaluation, which required the prototype to be evaluated against heuristics, it was decided to take advantage of regular meetings between the principal researcher and three other HCI experts, and to evaluate the prototypes using the ‘quick and dirty’ technique. This approach provided fast, informal feedback on these early prototype design ideas, which were subsequently integrated into the second version of the WOZ prototypes.

The goals for the evaluation of the second version of the WOZ prototypes in preliminary study two were (1) to ensure that the designs supported the usability requirements, and (2) to provide feedback to redesign the prototypes for experiment one. The two paradigms that were most suited to fulfil these goals were ‘usability testing’ and ‘field studies’. Although field study techniques have the advantage of providing ecologically valid data, they pose a particular problem when the system being evaluated is mobile, as it was in the work conducted for this thesis. For example, contextual inquiry (Beyer and Holtzblatt, 1998) was considered as a potentially useful field study technique, and involves interviewing users in their natural work environment. However, this is difficult in the case of voice-based mobile phone services, because there is no way for a user to pause their interaction if interrupted by the experimenter. The experimenter may therefore only watch the user, which would require participants to be observed for long periods of time, during which the experimenter would have to stay in the background making detailed written notes, and interrupting only when events arose which were thought to relate to the focus of the research. A further problem with contextual inquiry is that it requires substantial support from experienced contextual inquiry practitioners, which was not available for the work conducted for this thesis. Finally, it provides qualitative data, which would make it difficult to assess whether the usability goals had been achieved. The usability testing paradigm was therefore selected as being most appropriate, and consisted of the user testing study (preliminary study two), described fully in chapter

four. User testing is an applied form of experimentation used by designers to test whether the product they develop is usable by the intended user population to achieve their tasks (Dumas and Redish, 1999). The aim is to gather information about users' performance with the system, the analysis of which may be helped and supported by observational data, answers to user-satisfaction questionnaires and interviews, and data provided by monitoring software. User testing and experiments both measure performance but may be contrasted by their aims. Whereas experiments aim to discover new knowledge by hypothesis testing, user testing aims to gather data to inform and improve the design of successive prototypes. User testing needs to be carefully planned, and it should be possible to repeat the tests to obtain similar results, but the results are not required to be exactly replicable. The number of users recommended for user testing is typically 5 to 12, and the results are usually presented as means and standard deviations (Dumas and Redish, 1999). Monitoring software, satisfaction questionnaires, and interviews were used to collect data as part of the user testing for the work reported in this thesis. These techniques were also used for data collection for the three experiments, and will therefore be discussed in the second part of this chapter (see section 3.2.2.5).

3.2.2 Methodology Two: Experimental methodology

An experiment is one of the most powerful methods of evaluating a design or an aspect of a design (Dix et al. 2004). The aim of an experiment is to test a hypothesis that predicts a relationship between two or more events, known as variables. The hypotheses are tested by manipulating one, or some, of the variables (Preece et al. 2002). According to Robson (2002), an experiment involves: the assignment of participants to different conditions; manipulation of one or more of the independent variables; the measurement of the effects of this manipulation on one or more of the dependent variables; and the control of all other variables. For the work reported in this thesis, the experimental method allowed different interface designs to be tested under controlled conditions, and for the results to be statistically analysed for significance. There are two distinct types of experiment: those that are performed in the laboratory and those that are conducted in the work environment, or 'in the field'.

When an experiment is conducted in a laboratory the participant must be taken out of the environment in which they would normally use the system and situated in the

controlled environment of the usability laboratory. An advantage of the laboratory is that it allows the isolation and control of variables, in order to accurately measure cause and effect (Coolican 1990), thus allowing different designs to be compared. In addition, the laboratory is stocked with the technology and apparatus to allow extensive data recordings, and offers the participant an environment free of everyday distractions. The laboratory also offers a situation to test systems that may be located in dangerous or remote locations, and can also be effective at testing constrained single user tasks, such as Internet usage. However, there are a number of problems associated with this approach. Firstly there is the lack of realism, which Aronson and Carlsmith (1986) divide into both experimental realism and mundane realism. Experimental realism refers to the extent to which the experiment presents the participant with a realistic situation that has an impact on them, as would a realistic event. Mundane realism refers to the degree that the experiment presents the participant with an event that is likely to occur in the real world. The danger with running an experiment that lacks both types of realism is that the experimenter is simply recording a situation that never occurs in the real world, and therefore lacks ecological validity.

Coolican (1990) has isolated another two potential weaknesses: artificiality; and the inability to generalise. By artificiality it is meant the way in which the contrived situation created by the laboratory setting affects the participant. They may feel anxious or overawed by the laboratory setting, feelings which can be compounded if the experimenter sticks too rigidly to standardised protocol, and neglects the normal human interaction norms, leading to a negative impact on performance. Bias may also occur as a result of the demand characteristics of the experimental situation, which means that the participant may alter their behaviour according to their interpretation of what the experiment is testing and what the experimenter requires of them, an effect that has been shown to be most pronounced amongst participants who have volunteered for an experiment (Rosenthal and Rosnow, 1975). Although such biases may be mediated by keeping experimenter-participant interactions to a minimum, which is often the case with human-computer interaction experiments, many have argued that these weaknesses lead to results, which cannot be generalised to the real world beyond the laboratory.

The alternative to laboratory studies is the use of field studies, which situates the participant in their natural real world environment, and allows the experimenter to capture interactions between systems, and other people, that would not have occurred in the laboratory (Coolican 1990). In field studies, the participant interacts in real world conditions of ambient noise, movement, interruptions, and distractions, which are hard to replicate in the laboratory and which enables results to be generalized to the real world, thus promoting external validity. The natural situation of the field experiment reduces the demand characteristics of the experiment through the use of both experimental and mundane realism, and therefore reduces the tendency for participant biases to affect performance. Robson (2002) states that, if an ethical means of random allocation of participants to experimental conditions can be achieved, then a field study is preferable to a laboratory study. It is for this reason that, for the three experiments reported in chapters five, six, and seven of this thesis, field experiments rather than laboratory experiments were conducted.

3.2.2.1 The experimental plan

When planning an experiment it is necessary to consider the purpose of the experiment in terms of the variables that will be manipulated and measured, the sample of participants that will perform the experimental tasks, the design of the experiment, and finally the statistical tests that will be used to analyse the data (Coolican, 1990). These four components will be considered in the following sections.

3.2.2.2 Variables

A variable is any characteristic that can vary across people, or situations that can be of different levels or types (Breakwell, Hammond, Fife-Schaw 2000). A potential problem with carrying out field experiments is the lack of control the experimenter has over the environment, and therefore over extraneous, or confounding, variables, which can mask the effects on the variable that is of interest (Robson 2002). This can reduce the degree to which a field experiment can be accurately replicated. For the work reported in this thesis, efforts were made to balance the effects of confounding variables by both counterbalancing conditions and randomising the allocation of participants to conditions. In order for an experimenter to remain objective and to provide a valid method for measuring some part of a hypothetical construct, it is necessary to state an operational definition of the variables (Coolican 1990). This

provides a statement of what is being used as a measure for the construct of interest, for example, performance as a dependent variable, or metaphor type as an independent variable.

The international standard ISO 9241-11 (1998) defines usability as the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use. This refers to whether the system is effective in allowing the user to complete a specific task, whether the system is efficient in allowing them to do this, and finally whether the users actually like the system. Participants' subjective attitudes towards the system usability were used to measure satisfaction, whilst objective measures of performance were used to measure the effectiveness and efficiency of the system. The independent variables manipulated within each of the three experiments reported for this thesis are covered in experimental chapters five, six, and seven. Operational definitions for the subjective attitude variables, and the objective performance variables will be stated in the following section.

3.2.2.2.1 Operational definitions for the subjective variables

Subjective satisfaction is one of the five human factors goals proposed by Shneiderman (1987) for quantifying the efficiency and usability of an interactive system. With reference to speech recognition systems, Jones et al. (1989) recommend that an ergonomic evaluation should include measures of user attitudes. For the usability evaluation of voice applications, Tate, Webster and Weeks (1993) recommend the use of questionnaire and interview techniques to measure subjective satisfaction. Therefore, a consistent theme in these measures is the importance of measuring the users' subjective attitudes. The measures of subjective attitude that were decided upon were the factors important in the assessment of user perceptions of speech systems, identified by Hone and Graham (2000).

In their study, Hone and Graham (2000) aimed to develop a valid, reliable and sensitive measure of users' subjective experiences with speech recognition systems. The result was a questionnaire for the Subjective Assessment of Speech System Interfaces (SASSI). As a starting point they reviewed subjective evaluation measures for generic interaction systems research, including Shneiderman's Questionnaire for

Interaction Satisfaction ('QUIS' Shneiderman, 1998) and the Software Usability Measurement Inventory ('SUMI' Kirakowski, 1996). They concluded that these questionnaires failed to address features unique to speech recognition interfaces, such as the intuitiveness and habitability of the conversation.

Their next step was to review the usability evaluation methods specific to speech systems. They reviewed unstructured methods, such as the use of open interviews and overall rating scales, used by Nelson (1986), and Brems, Rabin and Waggett (1995) respectively. They found that the type of data elicited was difficult to translate into design requirements.

Next, they reviewed more structured methods, including the use of adjective pairs within rating scales (for example, Dintruff, Grice and Wang, 1985; Casali, Williges and Dryden 1990), and the use of a questionnaire format (for example, Zajicek, 1990) using a scale developed by Poulson (1987) to rate the perceived quality of software interfaces. Their final review was of the subjective attitude questionnaires developed by researchers at the University of Edinburgh to evaluate phone-based services. Questionnaires used in studies by Love (1997) and Dutton et al. (1999) used five or seven point Likert style rating scales (Likert, 1932), which required participants to respond to a number of statements, such as 'I found the system easy to use'. An overall mean was then calculated to provide a single measure of subjective satisfaction.

Hone and Graham (2000) concluded that most of the methods reviewed for the evaluation of speech systems suffered from four important weaknesses. Firstly, the items chosen for the questionnaires appear to have been based on the researcher's intuition, rather than on theory or empirical findings. Secondly, the questionnaires were not validated against other subjective measures and cannot therefore claim to measure what they were designed to measure. Thirdly, the reliability of the questionnaires, as measured by the test-retest reliability and the internal consistency, were not established. Finally, by taking an overall mean of the response items, an assumption was being made that all of the items were measuring the same underlying construct, which may not be the case.

In order to overcome the weaknesses associated with existing measures, Hone and Graham (2000) adopted a factor analytic approach. They pooled a total of 50 statements derived from existing questionnaires, and from a review of relevant literature, to form a Likert scale questionnaire, which they used to evaluate four different speech applications. They then conducted an Exploratory Factor Analysis on the data from the completed questionnaires, which suggested the presence of 6 factors, which are defined below, and which were used for the work conducted for this thesis:

1. **System response accuracy:** refers to the user's perception of the system as doing what they expect it to.
2. **Likeability:** refers to the user's ratings of the system as useful, pleasant and friendly.
3. **Cognitive demand:** refers to the perceived amount of effort needed to interact with the system and the feelings resulting from this effort.
4. **Annoyance:** refers to the extent to which users rate the system as repetitive, boring, irritating, and frustrating.
5. **Habitability:** refers to the extent to which the user knows what to do and knows what the system is doing.
6. **Speed:** refers to how quickly the system responds to user inputs.

3.2.2.2.2 Operational definitions for the objective variables

A review was conducted of: the performance measures for interactive systems proposed by Shneiderman (1987), Wixon and Wilson (1997), Preece et al. (2002), and Dix et al. (2004); measures of efficiency for hypertext information retrieval (Mohageg, 1992); and performance measures for speech systems evaluation (Dutton et al. 1999). From this review four main categories of performance measures emerged: time taken to complete a task, successful task completion, error rate, and navigation path through the system. Table 3.3 shows how these four categories map onto effectiveness and efficiency, which are the two facets of performance-based usability defined by the international standard ISO 9241-11 (1998). Table 3.3 also shows how the final eight objective performance measures, that were used for the experiments conducted for this thesis, map onto the four categories of performance derived from the review.

Table 3.3. The 8 performance measures recorded for experiments one, two, and three

Usability performance goals from ISO 9241-11	Usability performance categories	Usability performance measures
Effectiveness	Task completion	Successful task completion
	Error rate	Number of 'no user response' Number of 'Repeat'
Efficiency	Navigation path	Number of nodes used Number of nodes used as a percentage of the optimum
		Number of 'Return'
	Time	Time taken to complete the task as a percentage of the total prompt time Number of prompt interrupts as a percentage of the total number of nodes

An explanation of each of the 8 objective measures of task performance that were collected during participant interactions is given below.

1. Time taken to complete tasks (as a percentage of the total prompt time): The time taken to complete each task was logged by monitoring software, starting when the first dialogue of level 1 was played (entry point), and finishing at the end of the 'Exit' dialogue. The absolute time taken to complete tasks could not be compared between groups because the prompts were of different lengths for each service. In order to be able to compare groups it was therefore necessary to use the time to complete a task as a percentage of the total prompt time. A score of 100% would indicate an instantaneous response (the response time and prompt time were the same). A score of less than 100% would indicate that the participant had interrupted the prompt before it had finished playing, and a score of more than 100% would indicate that the participant listened to the whole prompt and then waited before making a response. A lower score was taken as being an indicator of good user performance.

2. Task completion: The single most important feature of any automated phone service is that users can successfully complete their tasks with it. The percentage correct task completion was therefore an important objective measurement for this experiment. A higher score was taken as being an indicator of good user performance.

3. Prompt interrupts (as a percentage of the total number of nodes used): The extent to which participants interrupted prompts with a response may be an indicator of how well they learned to use the service, the ease with which they could predict options, and their ability to extend their experience from the real world to the service domain. The measure 'prompt interrupt' records the total instances of prompts being interrupted, but does not provide information about the relative frequency with which prompts were interrupted. For example, one participant may complete a task using 20 nodes, and another participant may complete a task using 10 nodes, with both using 10 prompt interrupts. Their frequencies of use, relative to the total number of dialogue nodes accessed, would therefore be very different. The first participant would have used prompt interrupts for 50% of the nodes, and the second participant would have used prompt interrupts for 100% of the nodes. It was therefore decided to calculate prompt interrupts as a percentage of the total nodes, rather than as an absolute measure. A higher score was taken as being an indicator of good user performance.

4. Total number of nodes used to successfully complete tasks: Each node within the service structure represents a speech prompt, which, when selected, provides a user with spoken menu options allowing them to progress through the service. The number of nodes selected by a participant in order to complete the task is therefore an indicator of the efficiency of the service. A lower score was taken as being an indicator of good user performance. This measure allows a comparison between users on a task, but not between tasks of different levels of difficulty, and therefore measure 5, which is explained below, was also used.

5. Number of nodes used to complete tasks (as a percentage of the optimum): the minimum number of nodes to complete a task (optimum number) was calculated for each task, and the actual number of nodes taken by a participant was then calculated as a percentage of this. If a task was completed using the optimum number of nodes, then the minimum score of 100% was achieved. A score of more than 100% means that the participant used more than the optimum number of nodes. This measure allows deviations from the optimum path to be analysed. A lower score was taken as being an indicator of good user performance

6. No user response: This was a measure of a participant's failure to respond to system prompts after a time out period of five seconds. Instances of no response after five seconds were assumed to indicate that the participant was either lost within the service, unable to remember the appropriate menu options, or failed to understand the message or prompt. A lower score was taken as being an indicator of good user performance

7. Total number of times 'Return' function used: The total number of times the user returned to the first dialogue prompt of the service from levels 3, 4, or 5 of the service structure. A return to level 1 from level 2 would just be the equivalent of using the 'Back' command. This measure was only recorded if the use of the 'Return' function resulted in fewer overall nodes being used to complete the task than if the 'Back' function was used instead. In such an instance, use of the return function would represent a more efficient interaction with the service. A lower score was taken as being an indicator of good user performance

8. Total number of times 'Repeat' function used: The total number of times the user requested a repeat of a dialogue prompt. If users requested more than 3 repeats, they were automatically transferred to the human operator. A lower score was taken as being an indicator of good user performance

3.2.2.3 Design

In order to produce reliable and generalizable results, an experiment must be carefully designed. There are two main types of experimental design: between subjects and within subjects. In a between subjects design, each participant is assigned to a different condition, and there are at least two conditions. In the experimental condition the experimenter manipulates a variable, whereas in the control condition the experimenter keeps the variable constant. The same measures are taken for both conditions and, in this way, any changes to the experimental condition can be attributed to the variable manipulated. As part of the work conducted for this thesis, an example of a between subjects variable that was used for experiments one, two, and three, was the type of service that participants used. In the control condition participants used a non-metaphor service, whilst in the experimental condition participants used a metaphor-based service. The advantage of this design is that,

because each participant performs in only one condition, learning effects are controlled. A disadvantage is that more participants are needed, and that differences between groups and between users can negatively affect the results (Dix et al. 2004). For the work reported in this thesis, matching all groups for age, gender, and computing experience mediated these negative effects.

When using a within-subjects design each participant performs under each condition, and therefore each condition is automatically matched for individual differences such as age, gender and personality. As part of the work conducted for this thesis, an example of a within-subjects variable was the experimental trial, with participants performing tasks with a specific service over a number of trials. When a participant performs in more than one condition, an order effect is introduced. This order effect can be partially addressed through the use of a counterbalancing strategy (Robson 2002), whereby participants are randomly assigned to different orders of conditions so that an equal number of participants perform each of the possible order of conditions. However, for the work reported in this thesis, participants using different services performed the same tasks within the same number of trials, which enabled learning effects to be compared across groups. It was not therefore necessary to balance out order effects.

In experiments, such as those reported for this thesis, which have more than one independent variable, a mixture of the two designs can be used to produce a mixed factorial design. In such designs, one independent variable is between-subjects, and another is within-subjects, which permits the assessment of each independent variable as well as the interaction between the independent variables (Dancey and Reidy, 2002). The independent variables and the number of levels within each are covered in more detail in the individual experimental chapters five, six, and seven.

3.2.2.4 Sample of participants

One of the main aims of conducting an experiment in HCI is to be able to generalise the results to the intended user population. It is therefore vital that a sample of participants is chosen for the experiment that is representative of the intended target population. Population refers to all the cases of people who will use a particular system, whereas a sample is a selection from the population. For the work reported in

this thesis, the target population were English speakers who owned a mobile phone, and who had some experience of using automated phone services. The requirement for participants to be English speakers was due to the service output being presented as spoken English, and the need for an assumption to be made about each participant's level of speech comprehension and vocabulary. In addition to this, the speech output was synthetic, which can be perceived as sounding unnatural and is more difficult to comprehend (Leedham, 1991). The requirement for participants to have used automated phone services was due to the need to recruit a sample of participants who currently use systems similar to the one being evaluated. Another reason for this requirement was to exclude complete novice users from the sample, which would have negatively skewed results. These participant requirements were met through the use of filter questions during participant recruitment.

Participants were recruited through a number of methods. Firstly, opportunistic sampling was used, whereby colleagues and friends were approached and asked if they would mind taking part in the experiment. Secondly, participants were recruited from a participant pool scheme at the University of Portsmouth. The participant pool consisted of undergraduate psychology students who were required to participate in a specific number of hours of experiments as part of the credits towards their degree. Thirdly, participants were recruited as the result of an email advertisement campaign at Brunel University. These participants were financially compensated for taking part in the experiments. The different recruitment methods may have influenced participant motivation levels. Therefore, whenever possible, experimental conditions were balanced for participants recruited from different sources, in order to balance out any motivation effects. The numbers of participants used, and the descriptive data describing them are provided in the methodology sections of the relevant experimental chapters of this thesis.

To ensure that all experiments were conducted in accordance with the ethical principles outlined in The British Psychological Society (1993) 'Code of Conduct, Ethical principles, & Guidelines', the experiments were submitted to the Department of Psychology Ethics Committee at the University of Portsmouth, who reviewed the proposed experiments. All experiments were designed based on the guidance from

this committee, and in light of the British Psychological Society guidelines, ethical approval for each experiment was obtained before the experiments were conducted.

3.2.2.5 Data collection instruments

Subjective attitude data was collected using a usability questionnaire (see Appendix 2), objective performance data was collected using monitoring software, and qualitative data was collected using interviews. These techniques, and the specific tools used for the work reported in this thesis will be discussed in the following sections. Questionnaires and tests were also used to collect information about participants' individual characteristics, specifically age, gender, previous mobile phone experience, previous fixed line telephone experience, previous computing experience, verbal ability, spatial ability, cognitive style, working memory, and attitude towards mobile phone usage in public. The research instruments used to elicit this information will be described within the appropriate experimental chapters.

3.2.2.5.1 Usability questionnaire

User satisfaction is one of the three facets of usability defined by the ISO 9241-11 (1998). Satisfaction refers to how the user feels about the way that they interacted with the system, and can be determined by using a user satisfaction questionnaire. A questionnaire is a method for the elicitation, recording, and collecting of information about users' opinions of a system. The researcher should design the questions to help achieve the goals of the research and to answer the research questions (Robson, 2002). To gather subjective attitude data for the work conducted for this thesis, a 50-item, 7-point Likert scale usability questionnaire was designed. Likert scales are used for measuring opinions, attitudes, beliefs, and have been widely used for evaluating user satisfaction with products. Coolican (1990) proposed a number of advantages of using the Likert technique: it is more natural to complete and maintains the respondent's direct involvement; it has been shown to have a high degree of validity and reliability; and it has been shown to be effective at measuring changes over time. Scales usually range from 1 to 3 points, to a maximum of 1 to 9 points, but it is generally agreed that taking the middle ground, by using scales of 1 to 5, or 1 to 7, is the most effective method (Dix et al. 2004). For the work reported in this thesis, it was therefore decided to use a scale of 1 to 7, as follows:

Strongly agree	Agree	Slightly agree	Neutral	Slightly disagree	Disagree	Strongly disagree

It is important to consistently label the scales, so that, for example, a ‘1’ always indicates low agreement, whilst a ‘7’ always indicates high agreement. A final, but important, point is the need to run a pre-test with a few colleagues or friends to gather feedback on aspects of the questionnaire, such as whether the instructions and questions are clearly worded, unambiguous, and avoid using jargon.

The questionnaire was designed in a systematic way. Firstly, a wide range of questionnaire items were generated from the work conducted as part of the SASSI project (Hone and Graham 2000), and the work conducted on automated telephone service evaluation at the University of Edinburgh in collaboration with British Telecom (Love, 1997; Foster, McInnes, Jack, Love, Dutton, Nairn, and White, 1998; McInnes, Nairn, Attwater and Jack, 1999; Dutton et al., 1999). These items were matched to one of the six subjective factors identified by (Hone and Graham 2000). This process was iterative, and involved removing some items that were deemed too similar, and then assigning the remaining items to one of the six factors.

The content validity of the resulting items was then reviewed to determine the extent to which the test actually measures what it is supposed to measure (Rust and Golombok, 1989). In terms of validity, the use of the Likert scale questionnaire has been proven to be a valid technique for gathering attitude data. However, the degree to which the questionnaire was measuring what it was designed to measure was evaluated through a process of content validity. This process involved asking colleagues with an expert knowledge of the domain to evaluate the content of the questionnaire to ensure that the items were representative of the area that they were supposed to cover, and were not weighted towards specific aspects of the area. This process was conducted with three HCI experts, and was consequently revised to ensure that each factor was measuring what it was supposed to measure. This process resulted in a total of 50 questionnaire items matched to the six factors. These items were then divided into two groups of equal numbers of positively and negatively worded statements, in order to prevent bias effects caused by a respondents tendency

to habitually agree or disagree with the statements, rather than providing responses that actually mirror their attitudes. The items representing each factor can be seen in Appendix 1, and the final questionnaire can be seen in Appendix 2. This questionnaire was used for all three experiments.

3.2.2.5.2 TrueActive monitoring software

Monitoring software allows interaction logging to be conducted. Interaction logging is an automated technique for gathering performance data that enables the researcher to gather large quantities of data during the experiment, and which frees them up to concentrate on other activities, such as prototype simulation, or observation. Interaction logging is enabled through the use of specialist monitoring software that can be configured to record user actions such as key presses and mouse movement, which can be time-stamped to provide a record of the length of each interaction event. For the work reported in this thesis, TrueActive monitoring software (TrueActive Corporation, Kennewick, WA) was used to log the message prompts that were requested by participants, and the amount of time spent listening to each prompt. By measuring these two aspects of users' interactions, all of the objective performance measures required to assess the effectiveness and efficiency of the service could subsequently be calculated. Monitoring software was used for the user testing conducted in preliminary study two, and for all three main experiments.

3.2.2.5.3 Interviews

Interviews can be thought of as 'a conversation with a purpose' (Kahn and Cannell, 1957) and provide a direct way of gathering information from users about their experience of using a system. Interviews may be classified according to the degree of control the interviewer imposes onto the conversation and a commonly made distinction is between structured, semi-structured, and unstructured. Structured interviews consist of pre-determined questions that are asked in a set order. Semi-structured interviews have pre-determined questions but the order is not set and the interviewee is prompted to expand on points they make until no new information is forthcoming. Unstructured interviews do not rely on a set of questions, and allow an informal conversation to develop within an area of interest.

The type of interview used will depend on the evaluation goals. The purpose of using interviewing for the work reported in this thesis was to gather information that could serve to clarify and illustrate the meaning of the quantitative data, and to allow the participants to fully express their opinions about the system. The semi-structured interviews used provided a degree of structure, but also enabled the participants to expand on their answers. This type of interview must also be broadly replicable and was therefore judged to be more suited to experimental testing, which requires that an experiment should generate replicable data. The interviews were conducted at the end of the testing sessions, and the actual questions used for experiments one, two, and three can be seen in Appendices 14, 15, and 18 respectively.

3.2.2.6 Data analysis

The statistical tests that will be applied to the data must be decided upon during the planning stage of the experiment to ensure that the data can be analysed and that this analysis will allow the hypothesis to be either supported or rejected (Breakwell et al., 2000). ANOVA is a powerful parametric means of analysing differences between three or more conditions, and was the technique used for the work reported in this thesis. A multivariate approach was adopted as several sources of variance were measured (dependent variables), which taken together comprised a composite measure of usability. Due to the presence of multiple dependent variables and more than one independent variable in each of the experiments, the multivariate analysis of variance MANOVA was the primary analysis method used, and is simply an extension of ANOVA.

Another technique used for predicting outcomes was multiple regression. Multiple regression allows a researcher to discover whether a number of variables, called predictor variables, can be used to predict the value of a single variable, called the criterion variable (Dancey and Reidy 2002). The technique is similar to correlation but whereas correlation analyses allow the strength of association between variables to be analysed, multiple regression allows conclusions to be made about how much a variable (criterion variable) will change if a number of other variables are changed (predictor variables). This technique was used to investigate the effects of individual differences on attitude and performance with the services in experiment two.

3.3 Chapter summary

This methodology chapter has described the general methodologies and techniques used for the work conducted on this thesis, provided a justification for their selection, and shown how the overall system design was planned according to the HCD process. The HCD process provided an iterative means of developing speech-based mobile phone service prototypes that were designed with high levels of user participation, in order to produce a realistic interface for subsequent metaphor evaluation. The experimental methodology then provided a controlled means of comparing the different prototype designs to evaluate which was the most usable. The following chapter reports the two preliminary studies that were conducted for this thesis. Preliminary study one was designed to generate categories of interface metaphors, and to select and describe specific metaphors that may be applicable to a speech-based mobile city guide service. Preliminary study two was a user testing study, designed to provide feedback that could be integrated into the redesigned mobile phone service prototypes implemented and evaluated for experiment one, and reported in chapter five of this thesis.

:: CHAPTER 4

Designing interface metaphors for automated mobile phone services

4.1 Introduction

The two preliminary studies conducted as part of this thesis are reported in this chapter, and represent the design and evaluation stages of the HCD process that was followed to develop the metaphor-based prototype mobile phone services for the work reported in this thesis. These prototypes were then tested and evaluated under controlled conditions in experiments one, two, and three. The first preliminary study was an exploratory study conducted to evaluate the potential of different techniques for generating, selecting, and developing interface metaphors for speech-based mobile phone services. The planned output was a shortlist of the metaphors that were considered by participants to be most applicable to speech-based mobile phone services. The second preliminary study was used to develop a sub-set of these metaphors into metaphor-based prototypes of a mobile city guide service, and then to evaluate them through a process of user testing. Important feedback was generated from participants during user testing to improve the design of the prototype services compared in experiment one.

4.2 Preliminary Study 1: Generating, selecting, and developing interface metaphors

4.2.1 Introduction

A point established as a result of the literature review was that interface metaphors have not been successfully applied to the design of speech-based automated mobile phone services. The exploratory study reported here aimed to address this discrepancy by extending previous methodologies and models from GUI design to the design and development of metaphor-based phone services. Through a HCD process, metaphors were generated, selected, developed and utilised by participants to explain how to perform tasks with two different automated phone services: a telephone internet service (TIS), and a telephone city guide service (TCGS). The primary objectives of this study were:

1. To compare the usability and productivity of card sorting and sketching as methodologies for selecting and developing interface metaphors for speech-based automated mobile phone services
2. To formulate a model of metaphor categories that may be applicable to speech-based automated mobile phone services
3. To select the three most suitable metaphors that could be used to implement prototype versions of a speech-based automated mobile city guide service for preliminary study two.

The following sections discuss brainstorming, card sorting and sketching as methodologies for generating and selecting interface metaphors, and the reasons for their use in the current study. The POPITS model, which was used to develop the metaphors selected, was previously discussed in section 2.3.4.

4.2.1.1 Generating interface metaphors - brainstorming

Alty et al. (2000) proposed a 6-step framework for the generation and description of metaphors (see section 2.3.4), which recommended brainstorming as a useful technique for initial elicitation. Brainstorming can be used early in the design process of a new and innovative system as a means of generating creative design ideas (Osborn, 1963; Jones, 1980; de Bono, 1992). The technique has been widely used

within design, and involves a group of task experts coming together to focus on a problem. The key features of brainstorming are the fast generation of many new ideas, and the absence of analysis and judgemental evaluation. This enables everyone in the group to gain a better understanding of the problem space. Alty et al. (2000) suggest writing the system functionality on a board, selecting related items of functionality, and then mapping real world processes onto them. According to Palmquist (2001), however, users are not good at generating their own metaphors, but can choose and articulate metaphors from a selection provided. It was therefore decided that, for the current study, the principal researcher and three other HCI experts would use brainstorming to generate a list of potential metaphors, whilst retaining the possibility for participants to develop their own.

4.2.1.2 Selecting interface metaphors – card sorting and sketching

In order to help participants to choose the best metaphors for the service, and to adhere to a HCD process, it was necessary to provide suitable techniques to allow participants to represent the way in which they naturally conceive of mobile phone services. Rather than giving the participants a diagram of the structure of a service, and thereby imposing a model onto them, the participants were asked to create a structure themselves by means of both card sorting and sketching. Both methodologies have been used within HCI for simple prototyping during the development of the design model, but in this case were used to develop the participants' conceptual, metaphor-based model.

Card sorting is a HCD method for discovering the latent structure in an unsorted list of statements or ideas (McDonald and Schvaneveldt, 1988). Card sorting allows the designer to explore how people group items, so that structures can be developed that reflect the structure in which the users expect the ideas or concepts to be presented, thereby maximizing the probability of users being able to find items. According to Rosenfeld and Morville (1998), card sorting can provide an insight into users' mental models, by providing a visual representation of the way that they organise information and conceive of information spaces within their own heads. Card sorting is therefore a way to elicit users' mental models of how they expect to find content or functionality. The investigator writes each menu item on a small card and asks the participant to sort

the cards into groups or clusters. The results of individual sorts can be combined and may be statistically analysed.

Robertson (2001) highlights some advantages and disadvantages of the technique. The advantages are that: it is quick, easy and cheap to conduct; it enables the designer to understand how users are likely to group items based on real people rather than intuition; and finally it identifies items that users may find difficult to categorize. The disadvantages are that: it is content-based and does not consider users' tasks, and the designer must therefore conduct a task analysis to ensure that the resulting information structure allows users to achieve tasks; it may capture superficial surface characteristics only, with participants not considering what the content is about or how they would use it to complete a task. For the work reported in this thesis, this problem was addressed by asking participants to provide a verbal description of how to complete a task using the structure produced as a result of their card sort.

Sketching is another HCD technique that can be used to explore design ideas, and can also be used to help designers to think about an organising metaphor for a system (Preece et al., 1994). The use of visual representations to discover creative solutions has been proposed as a fundamental mechanism of scientific discovery (Dreistadt 1968; Gooding 1996; Nersessian 1995; Qin and Simon 1995) as well as in other areas (Koestler 1964b; Shepard 1978; Johnson-Laird 1988). The strength of sketching is that it has been shown to be a valid technique to represent spatial mental models (Billinghurst and Weghorst, 1995), and it may therefore enable participants to provide a visual representation of their mental model of the spatial structure of items within an automated phone service. However, a potential disadvantage may occur in instances when users have very poor drawing skills, although this may be overcome in follow up sessions, in which the experimenter asks for clarification about any unclear aspects of the sketch. Sketching was therefore used for the work reported in this thesis to allow users to represent their conceptions of the service structure.

Card sorting and sketching yield an overall spatial structure for the phone service that will form a reference point to facilitate the choice of metaphors that share a similar overall structure with it.

4.2.2 Methodology

A two-condition within subjects design was used with the methodology type as the independent variable. All participants used both the card sorting and the sketching methodologies. Eighteen participants took part, and comprised 10 females and 8 males with ages ranging from 18 to 39 (mean = 23 years).

The study was conducted over nine separate experimental sessions lasting for one hour each, with two participants taking part in each session. Each session was split into two parts, both of which were video recorded, and observed by the experimenter from behind a two-way mirror in an observation suite. Participants were seated at tables opposite each other, but separated and obscured from each other by a large Velcro board.

Participants were given a brief explanation of automated telephone services, and then played a recording of someone performing a task using an automated telephone service (Unified Messaging Service). For part one of the experiment, participants were given a set of Velcro backed cards, on each of which was printed a menu option from a potential Telephone Internet Service (TIS), for example 'Shop'. The three top-level menu options from the TIS were: People, Shop, and Web Channels. The options were taken from the AOL online portal. The full range of menu options, and the actual service structure from which the items were taken, can be seen in Appendix 3.

Participants were then told that they had to arrange the cards on the Velcro board into a potential structure for the TIS. They were told that there was no right or wrong way to do this, and that they should simply arrange the cards into a structure that made sense to them. A time limit of ten minutes was imposed for this part of the study. The participants were then given a sheet containing a table of 60 'real world systems', which were the metaphors generated in the brainstorming session by the experimenter and the HCI experts (see Appendix 5). They were asked to choose and rank five of these systems, or to additionally generate metaphors of their own, that they considered to be most similar to the structure of the TIS they had produced. A rank of '1' indicated that the metaphor chosen had a structure most similar to the structure they had produced, whilst a rank of '5' indicated that the metaphor was the fifth most similar.

Participants were then required to take the metaphor they had ranked as being most applicable to the system structure, and to develop it according to the six areas of metaphor features proposed by the POPITS model. Finally participants were asked to explain to each other how to perform two tasks using the service they had constructed. The participants were asked to give these explanations using language derived from the metaphor they had ranked as ‘number 1’, and with reference to their final arrangement of cards. The tasks were:

- Explain how to shop for digital cameras
- Explain how to access travel destination guides

The rationale underlying these explanations was to evaluate whether participants were able to use the metaphors in their task explanations, and whether they incorporated features arising from their POPITS analysis into the explanations.

The second part of the experiment was similar to the first. The only differences were that menu options from a telephone city guide service (TCGS) were used instead of those from a TIS, and that participants were required to sketch a structure rather than sort cards. The top-level menu options from the TCGS were: Arts, Eating out, and Nightlife. The full range of menu options and the service structure for the TCGS can be seen in Appendix 4.

Data was collected in a number of ways. A ranking sheet was used to record the participants’ choice of their top five metaphors (Appendix 5). A POPITS sheet was used to gather data about the salient features of the metaphor that participants ranked as ‘number 1’ (Appendix 6). Video recording was used to record participants’ natural language explanations of their metaphor. Finally, a usability questionnaire was used to record the participants’ attitudes towards the usability of each experimental methodology (Appendix 7). The usability questionnaire used a seven point Likert scale, with a high score indicating a positive attitude towards the usability of the methodology. The five items on the questionnaire corresponded to the five usability goals that were considered to be most relevant when considering the usability of a

methodology, and were adapted from the usability goals for interactive systems proposed by Preece et al. (2002, p. 14).

4.2.3 Results

4.2.3.1 Usability of the methodologies

Paired samples t-tests were performed on the usability questionnaire scores to compare participants' evaluations of the two methodologies. There was a significant difference between the two methodologies for each aspect of usability tested (Table 4.1). The results indicated that card sorting was more usable than the sketching methodology for all 5 rated aspects of usability.

Table 4.1. Usability questionnaire results from preliminary study one

Usability	Card sorting		Sketching		t	df	Sig. (2-tailed)
	Mean	SD	Mean	SD			
Efficient	5.67	1.03	4.17	1.47	4.75	17	0-001
Easy to use	5.94	1.00	4.06	1.73	4.88	17	0-001
Easy to learn	6.39	0.85	5.22	1.40	3.97	17	0-001
Comfortable	5.50	1.04	4.33	1.61	2.81	17	0.01
Productive	5.78	1.06	3.50	1.62	6.03	17	0-001
Mean	5.86	0.82	4.26	1.32	5.81	17	0-001

4.2.3.2 Productivity of the methodologies

The data generated from the POPITS section was quantified. Each word or phrase that referred to a specific feature of the chosen metaphor was quantified as one feature. Table 4.2 demonstrates the coding system, showing one feature for each of the POPITS categories for the 'Shopping Centre' metaphor.

Table 4.2. POPITS features for the 'Shopping Centre' metaphor

POPITS feature	'Shopping Centre' metaphor
Properties	Automatic sliding doors
Operations	Walk between shops that specialise in different products
Phrases	Do you have this in a size XL?
Images	Window display
Types	Large open plan with different sub-sections
Sounds	Background music

Paired samples t-tests were then performed to compare the number of metaphor features generated using each methodology. Only one significant difference was found between the two methodologies, which was for the POPITS feature ‘Sounds’ ($t(17) = 2.38$; $p = 0.029$). There was no overall difference between the card sorting and sketching methodologies in their abilities to stimulate the productivity of POPITS features relevant to a metaphor (Table 4.3).

Table 4.3. POPITS results for the card sorting and sketching methodologies

POPITS feature	Card sorting		Sketching		t	df	Sig. (2-tailed)
	Mean	SD	Mean	SD			
Properties	3.67	2.43	3.06	2.13	1.83	17	0.09
Operations	2.33	1.53	2.22	1.63	0.33	17	0.74
Phrases	2.17	1.65	2.72	2.42	-0.97	17	0.34
Images	2.27	1.74	1.67	2.00	1.26	17	0.23
Types	1.61	1.42	1.72	1.84	-0.33	17	0.74
Sounds	1.44	1.25	0.78	0.88	2.38	17	0.03
Mean	2.25	1.09	2.03	1.22	1.13	17	0.27

4.2.3.3 Using metaphors to explain an automated telephone service

In order to investigate whether participants were capable of using their chosen metaphors to explain tasks, the explanations given were analysed and referenced against their digital photos and sketches. Three criteria were used to establish whether the metaphor had been used effectively (1) consistent use of language derived from the metaphor (2) whether the language used correlated with the prototype the participant had designed (3) whether the metaphorical description was comprehensible to the other participant.

On the basis of these criteria, and using a binary coding scheme, a ‘1’ was assigned to effective explanations, whilst a ‘0’ was assigned to ineffective explanations. For the card sorting methodology, 15 participants effectively used metaphor and 3 did not. For the sketching methodology, 10 participants effectively used metaphor and 8 did not. A McNemar test showed no significant change in the number of effective metaphorical explanations between methodologies ($p = 0.063$). This result suggests that the methodology used made no difference to the number of participants who were able to produce effective metaphorical explanations, and that the majority of participants were capable of producing effective metaphorical explanations. It must be noted, however, that participants were asked to give explanations involving language

that had the potential to make them feel self conscious or embarrassed, and were required to do this without being able to discuss and prepare their explanations beforehand with other participants. It is possible that this may have resulted in explanations that were shorter or less involved than they otherwise might have been.

4.2.3.4 Card sorts, sketches, and their explanations

The structures produced by participants using either card sorting or sketching were of different types, and the depth and detail with which metaphors were used in participants' explanations varied. In order to provide an indication of the types of structures produced by participants, and of the corresponding task explanations that matched the effectiveness criteria, a number of examples are provided (overleaf).

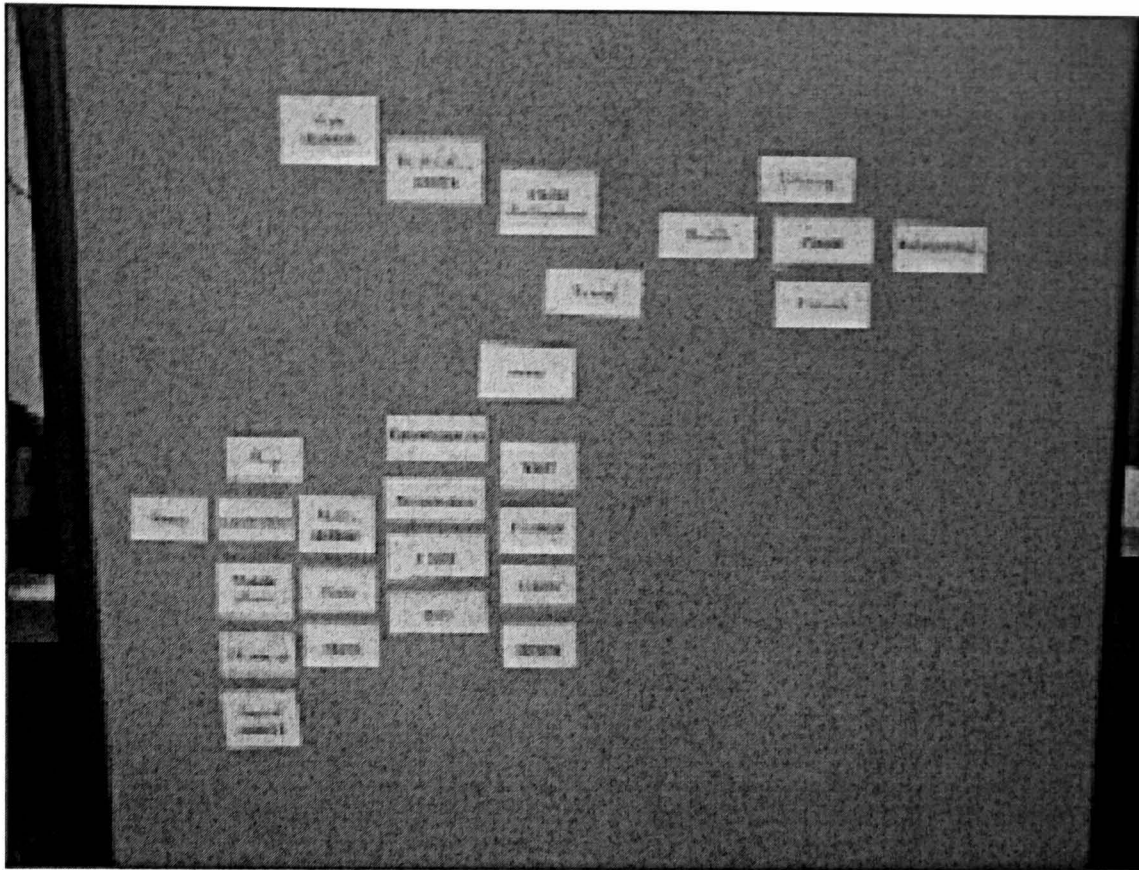


Figure 4.1. A card-sort of a road system metaphor

Road system: "...imagine you are on a motorway. These main headings are the main roads or lanes. You come to a junction or turning and you are thinking do I want to go to Scotland or do I want to go to Cornwall? Do I want to go to 'mens clothing' or do I want to go to 'electronics?' So I take the junction leading to 'electronics'...the further you go down the road you see the turn off signs for all of these things...so you are in Electronics Town and the next turning on the left is Digital Cameraville."

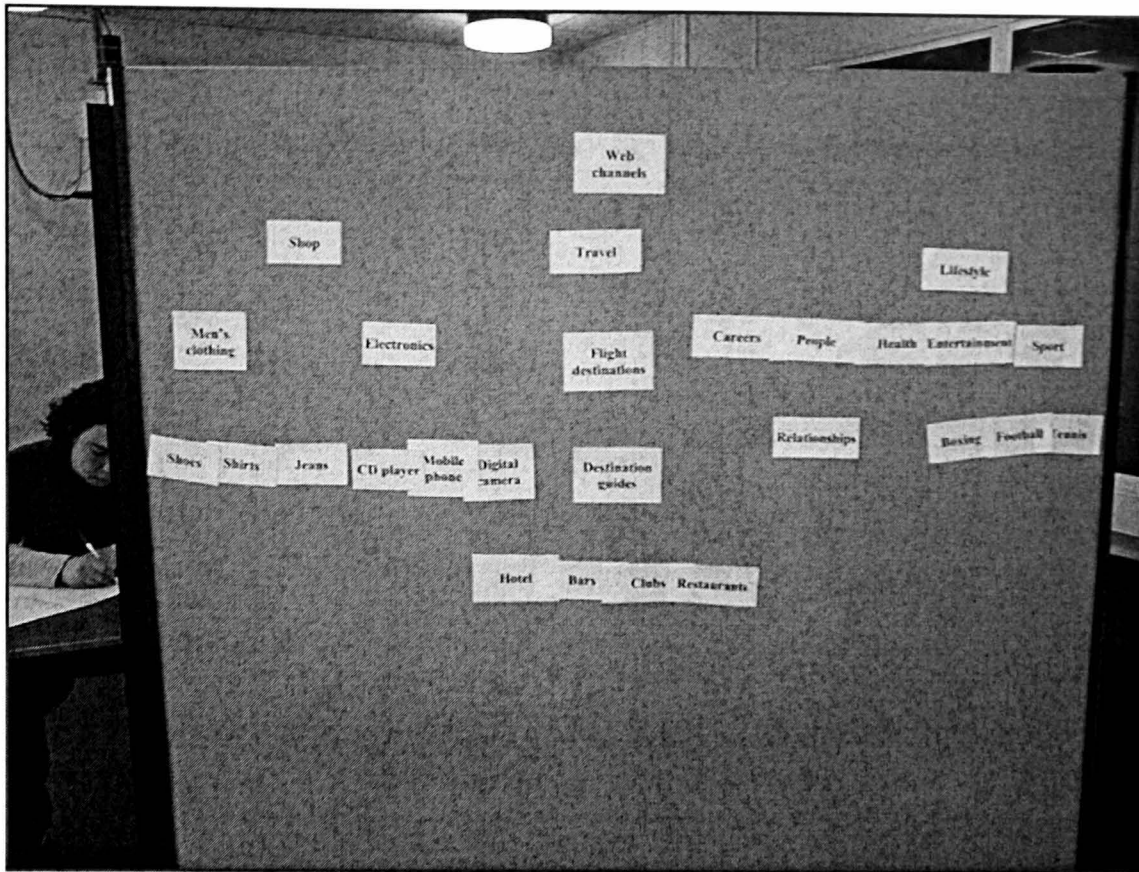


Figure 4.2. A card-sort of a filing cabinet metaphor

Filing cabinet: “You have a filing cabinet and there are three drawers. The top drawer is ‘shop’, the middle drawer is ‘travel’ and the bottom drawer is ‘lifestyle’. You need to go to the top drawer, and when you pull that open there will be two files in there. One is ‘mens clothing’ and the other is ‘electronics’... if you open up the ‘electronics’ file you’ll find three folders, one is a ‘CD player’, one is a ‘mobile phone’ and one is a ‘digital camera’, and you want the ‘digital camera’ one.”

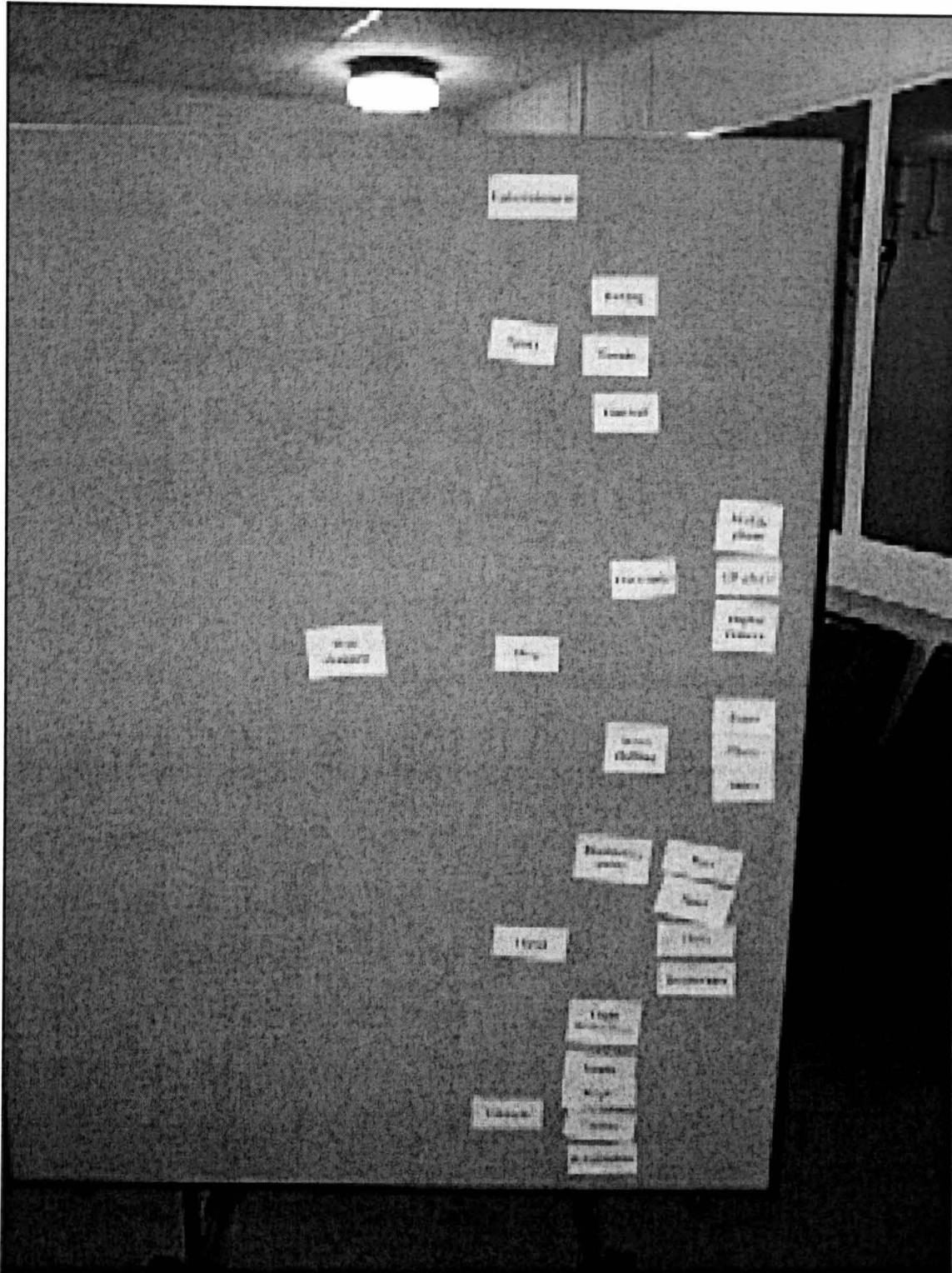


Figure 4.3. A card-sort of a human circulatory system metaphor

Human circulatory system: “My system was the circulatory system as in mainly starting off from a centre point say the heart...and then you branch off to the major veins and then these get smaller and smaller into more specific capillaries or whatever, veins. And then it’s just branching down to the smaller more specific points...and again that would branch off into two different veins say into ‘mens clothes’ and ‘electronics’...and the thing is also being able to go back round, that’s a big thing, so you should be able to go back round to the start and find the ‘travel’ vein and find the ‘travel destination guide.’”

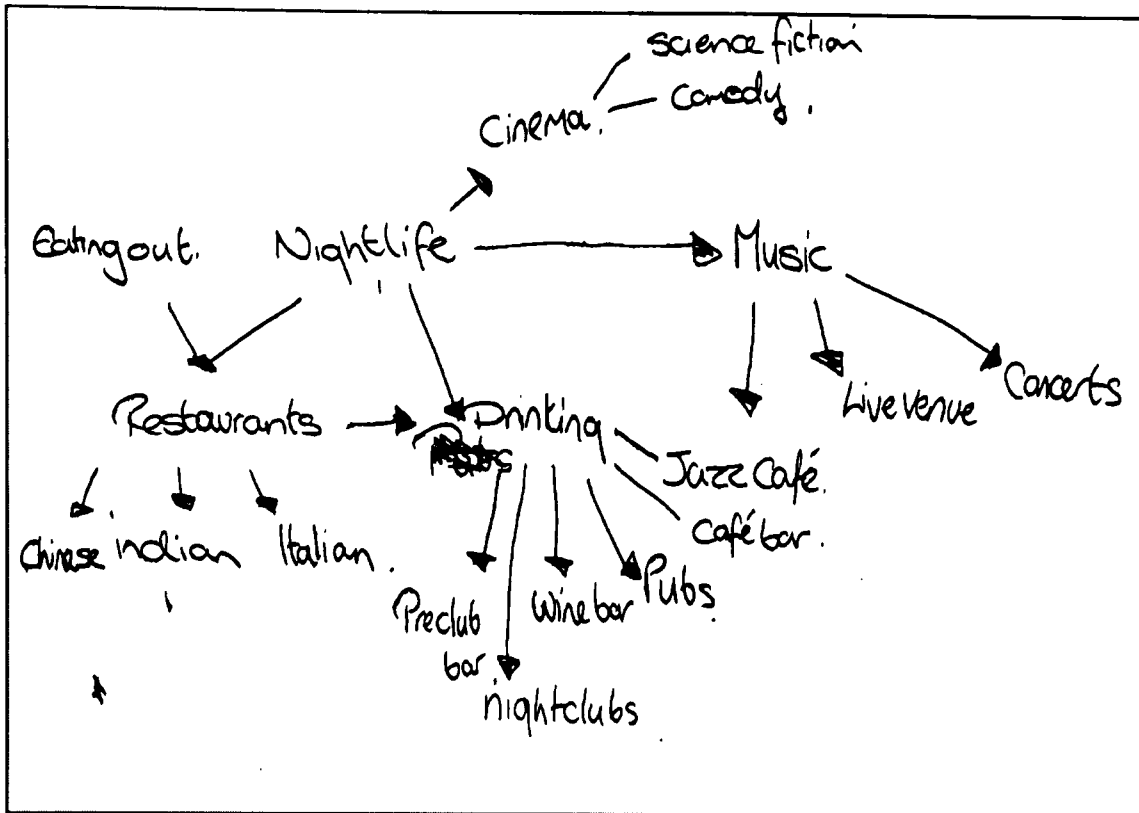


Figure 4.4. A sketch of a brain metaphor

The Brain: “Right I’ve chosen the brain because it is divided up into different areas, with each area having a different function, but it all links up and is somehow connected to allow a person to think. So if you start off with nightlife then that might make you think about drinking, and then with that chain of thought you could start thinking about pubs and nightclubs. Or if you are planning a night out, you might want to go to a restaurant, and then go on to a bar afterwards. So thinking in one area can stimulate thinking about another area, and it is all linked as a network of ideas.”

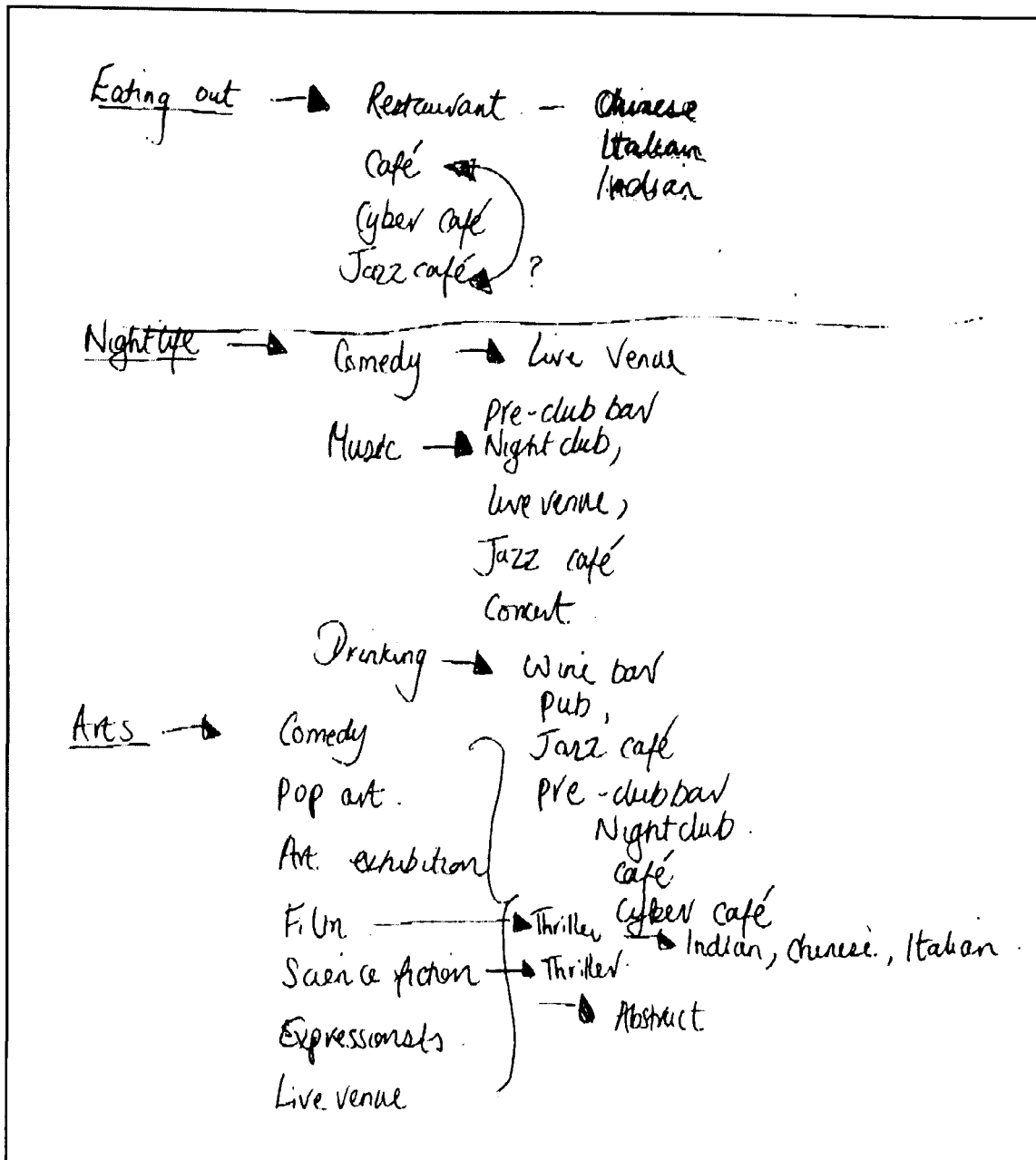


Figure 4.5. A sketch of a supermarket metaphor

Supermarket: “You know when you are in a supermarket and you are going shopping, and you’ve got different aisles. Different aisles are spread into bread, frozen, fruit, vegetables, so you go to different aisles and when you go to different aisles you’re gonna get your specifics. So basically if you wanted to go to an Italian restaurant, you know it’s a restaurant so you’d got to restaurants and that’s your section, say your aisle and then in your aisle you’d decide whether to get Chinese, Italian, or Thai and you’d go for Italian.”

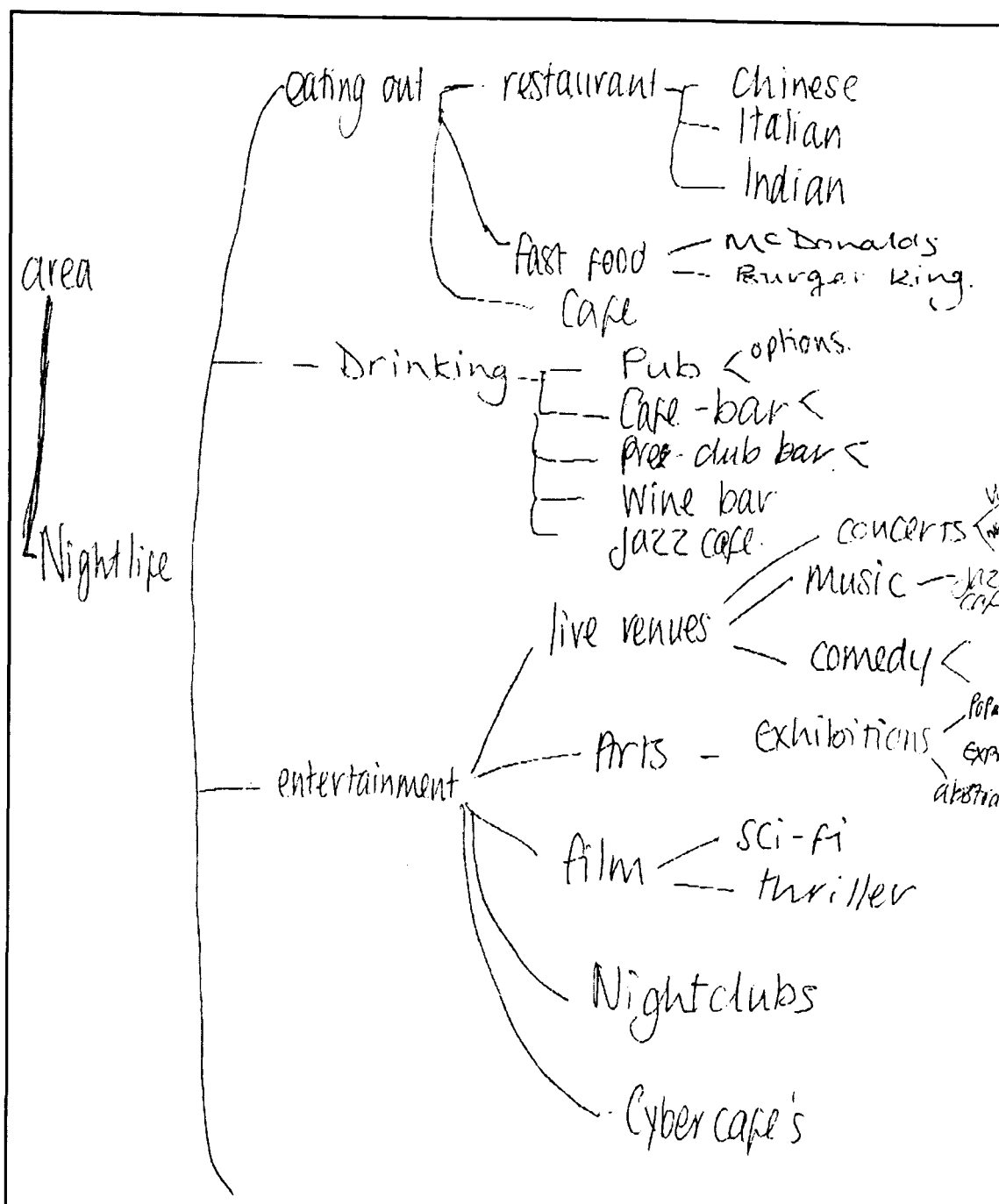


Figure 4.6. A sketch of a filing cabinet metaphor

Filing cabinet: “So my system is a filing cabinet and you’ve got three different drawers, eating out, drinking, and entertainment. To find out about restaurants you go into eating out, and in that drawer you’ve got a choice of restaurants, fast food and cafes kind of thing. You want an Indian restaurant so you’d go into restaurants and within that folder there’s three more folders so you go there and pick up the Indian restaurants. The bigger the system you might need to have separate filing cabinets to give you more information.”

4.2.3.5 POPITS features used in the explanations

Another measure of the level of detail used by participants in their metaphor-based explanations was the number of POPITS features they incorporated. Of the participants who gave effective metaphorical explanations, an association would be expected to exist between the number of POPITS features they generated, and the number of POPITS features they incorporated into their explanations. Bivariate correlations were therefore performed to measure the association between the number of POPITS features initially generated, and the number of POPITS features used within explanations, for each methodology. A significant positive correlation was found for participants using the card-sorting methodology ($r = 0.601$; $n = 18$; $p = 0.008$). There was no significant correlation found for the sketching methodology ($r = -0.019$; $n = 18$; $p = 0.941$). It may therefore be suggested that participants using the card sorting methodology were able to provide more detailed metaphorical explanations by incorporating a higher number of the POPITS features they had previously generated.

4.2.3.6 Metaphor categories

The ranking data from each telephone service was scored as follows: Rank 1 = 5 points, Rank 2 = 4 points, Rank 3 = 3 points, Rank 4 = 2 points, and Rank 5 = 1 point. The full list of metaphors and ranks can be seen in Appendix 8. The metaphors with the 10 highest mean rank scores were extracted, and represented the 10 most commonly occurring and highly rated metaphors. Table 4.4 lists the top 10 metaphors, the number of times each metaphor was selected, and the mean rank score.

Table 4.4. Top 10 metaphors for each telephone service

Top 10 position	Telephone Internet service				Telephone city guide service			
	Metaphor	No of times selected	Mean rank	SD	Metaphor	No of times selected	Mean rank	SD
1	Filing cabinet	9	1.72	2.08	Talking pages	6	1.44	2.12
2	Department store	8	1.44	1.89	Computer	6	1.11	1.78
3	Talking pages	6	1.28	1.90	Department store	5	1.06	1.86
4	Computer	6	1.28	2.02	The brain	5	1.06	1.95
5	Road system	6	1.17	1.92	Filing cabinet	6	1.00	1.71
6	Shopping	6	1.11	1.78	Road system	5	0.89	1.71
7	Supermarket	7	0.94	1.39	Transport	4	0.78	1.66
8	TV	3	0.50	1.29	Tree	3	0.67	1.57
9	Shopping centre	2	0.50	1.47	Supermarket	6	0.61	1.20
10	Transport	4	0.44	0.92	Circulatory system	3	0.50	1.20

Overall, 7 out of the top 10 metaphors were common to both services, suggesting that the same metaphors were considered to be applicable to different telephone services. The ‘Filing cabinet’ metaphor was ranked most highly for the TIS, and the ‘Talking pages’ metaphor for the TCGS. It was evident from examining the 10 metaphors that some of them were strongly related, shared similar structures and features, and evoked similar language. Examples include the Department store, Shopping, and Supermarket metaphors from the TIS, which all shared features such as aisles, shelves, and products. The Talking pages, Filing cabinet, and Computer metaphors from the TCGS all shared features such as files, directories, and hierarchies.

To investigate similarities between different metaphor types, the features and structural elements of the top 10 metaphors, which were generated by participants as a result of using the POPITS model, were analysed. This process was repeated for all of the top 10 metaphors, and was then extended to metaphors not ranked amongst the 10 highest. This process was iterative and involved assigning metaphors to a group that shared similar features, revised if unsuitable, and then reassigned to another group. Finally, five groups emerged that incorporated all of the top 10 metaphors, and which could be classed as metaphor categories. The 5 metaphor categories, example metaphors from each category, and the category definitions can be seen in Table 4.5.

Table 4.5. The 5 metaphor categories and their descriptions

Metaphor category	Example metaphors	Description
Hierarchical	Filing cabinet Talking pages Computer	The overall structure is a pyramid consisting of different levels of information, with each level being linked to the levels above and below it
Shopping venue	Department store Shopping centre Supermarket	A venue consisting of different sections and levels, with different means of travelling between them (walking, elevator etc), characterised by the division and sub division of goods into categories, and categories into shops
Transport system	Road system Bus route Transport	A complex network of routes connecting towns and cities. These routes are of different sizes and levels of importance, with different rules governing their use. The routes become more concentrated around urban hubs
Information provider	Television Website Internet	An information space with different channels and categories, with which it is possible to engage at different degrees of interactivity, and by using different input modes
Natural circular	The brain Circulatory system	A biological system separated into different functional areas, but connected by the process of circular flow

For each category, an overall mean score was calculated of the ranking points from each of the individual metaphors. It was then possible to determine the three highest scoring metaphor categories. From these three categories, the three most highly rated metaphors would then be selected to implement the metaphor-based service prototypes tested in preliminary study two. Table 4.6 shows the three highest scoring metaphor categories and the metaphor from each with the highest mean rank score.

Table 4.6. The 3 highest scoring metaphor categories and metaphors

Top 3 position	Metaphor category	Mean rank score	SD	Metaphor	Mean rank score	SD
1	Hierarchical	0.82	0.63	Filing cabinet	1.72	2.08
2	Shopping	0.74	0.65	Department store	1.44	1.89
3	Transport system	0.65	0.76	Road system	1.17	1.92

4.2.4 Discussion

The results from this study provide an important first step towards investigating the potential of interface metaphors for speech-based automated mobile phone services. The first research objective of this study was to compare the usability and productivity of card sorting and sketching as methodologies for facilitating the generation, selection, and development of interface metaphors for automated phone services. It can be concluded that the card sorting methodology was more usable. There was no overall difference between methodologies in their ability to stimulate metaphorical thought within the POPITS framework. However, participants using card sorting generated significantly more ‘Sound’ related features, which are particularly salient to the construction of speech-based telephone services.

The relationship between the number of POPITS features generated, and the number of POPITS features used in explanations for the card sorting methodology suggests that the card based structure produced by participants enabled them to initially generate POPITS features that were more relevant to the metaphor, and consequently easier to incorporate into coherent metaphorical explanations. Due to experimental constraints, participants were given a limited time period to produce an initial structure for each phone system. However, from observations made throughout the study, reactions to the ‘time up’ signal, and an informal interview given on

completion of the experiment, it was clear that the participants using the card sorting methodology were able to design faster, and more iteratively. Card sorting can therefore be recommended as an important visual means of stimulating metaphorical thought about telephone-based interfaces: structuring a telephone service using cards enabled participants to construct more meaningful associations between the structure and the potential metaphor. However, the experimental design used involved matching service type with methodology, which introduces a confound in the comparison of card sorting and sketching methodologies. The results must therefore be interpreted within the limitations of the experimental design.

The second objective was to investigate whether metaphors could be formulated into coherent categories. An analysis of the features and structures of the ten highest scoring metaphors suggested the presence of five categories of metaphor, which enabled the majority of metaphors that were chosen by participants to be classified. The following three individual metaphors, which were taken from the three highest scoring metaphor categories, may be most applicable to the mobile phone services examined: 'Filing cabinet'; 'Department store'; and 'Road system'. These metaphors were therefore selected to be used to implement metaphor-based versions of a speech-based mobile city guide service, which were user tested for preliminary study two, and which is described in the following section.

4.3 Preliminary Study 2: Improving the usability of mobile phone services using spatial interface metaphors

4.3.1 Introduction

In this study, a hierarchically structured, speech-based mobile city guide service was designed, which allowed users to access different categories of city-based information by listening to spoken service messages, and then to navigate to the required information by verbally selecting menu options. A standard version, using numbered menu prompts, and three different metaphor-based versions of the city guide service were evaluated. The three metaphors were 'Filing cabinet', 'Department store', and 'Road system', and were derived from preliminary study one. However, for the purposes of this study, the names of these metaphors have been changed to 'Office filing system', 'Shopping', and 'Travel system' respectively, in order to more

accurately reflect their features and structures. This study formed part of an iterative process of designing the dialogue and structure for the three metaphor-based services, and involved paper prototyping using flow diagrams, and WOZ-based prototyping, in conjunction with appropriate design guidelines. The three main objectives of the study were:

1. To investigate whether the use of different interface metaphors led to an improvement in usability compared to a non-metaphor version.
2. To investigate whether there were any differences in usability between the 3 metaphor-based services.
3. To generate feedback about the prototype services that could be fed into the redesign of the services that were compared in experiment one.

4.3.2 Prototype design and development

The service was designed as a hierarchical structure consisting of 5 levels of service messages and prompts. The user enters the service at level 1 (entry point), and navigates to level 5 to find the specific information they need in one of the 36 separate information containing nodes. The categories of information, and service menu options were designed on the basis of the user requirements gathered, which was reported in section 3.2.1.3. One of the 2 main branches of the hierarchy can be seen in Figure 4.7 below.

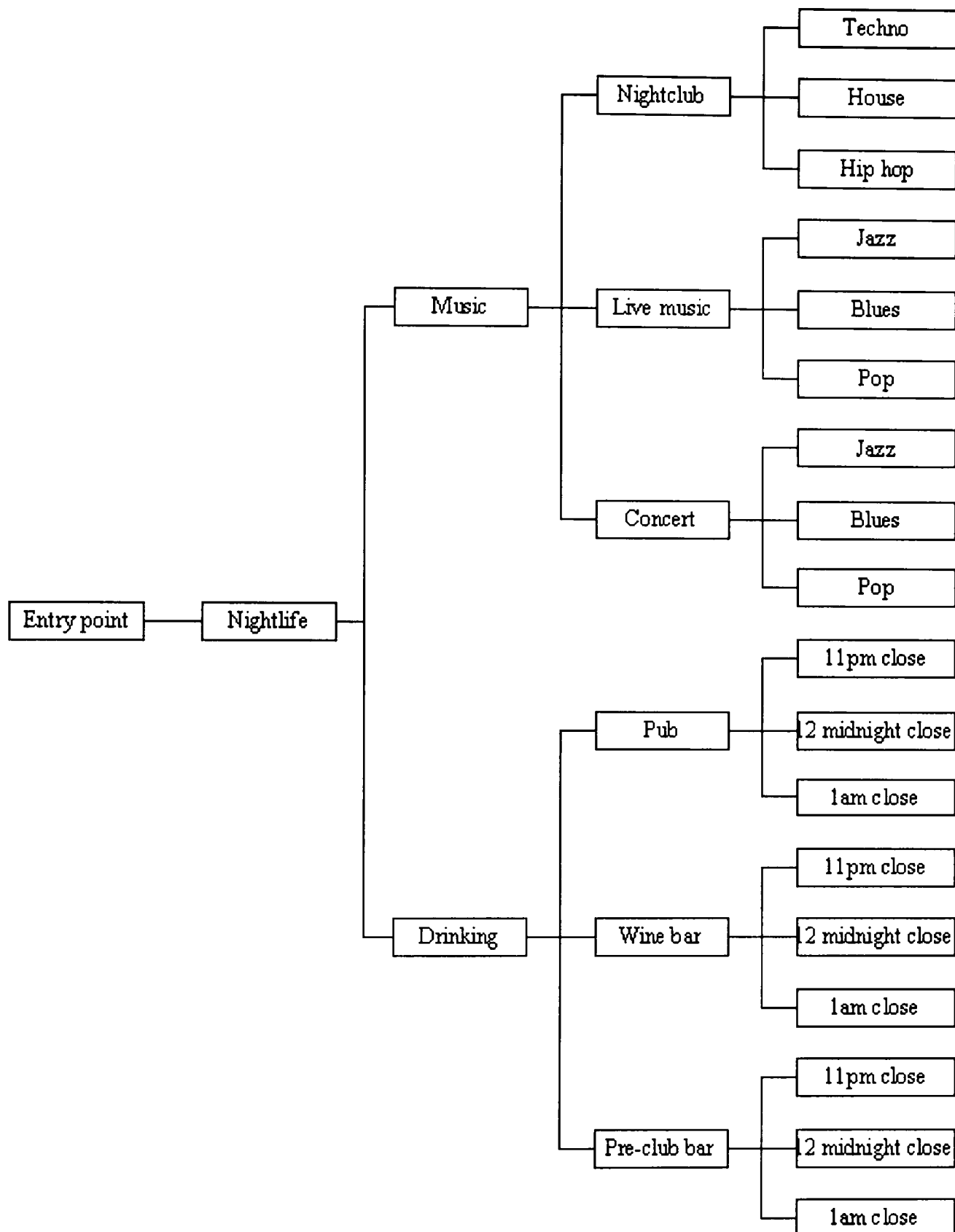


Figure 4.7. One of the 2 main branches of the city guide service menu hierarchy

One version of the service was designed in the same style as most commercially available automated telephone services. For example, ‘For option 1, say 1, for option 2, say 2...’ For the purposes of this thesis, this service will be referred to as the standard service. The designs of the other three versions of the service were based on three different metaphors derived from preliminary study one. These metaphors were

(i) a travel system (previously road system), in which participants were required to drive along different roads, take different turnings, and follow different coloured signs to find a tourist information office that could provide the information they wanted (ii) an office filing system (previously filing cabinet) in which participants were required to select from different drawers, sections, and coloured folders in a row of filing cabinets in order to find the information they wanted, that was located in a folder (iii) a shopping system (previously department store) in which participants could move around different floors, departments, and sections of a department store to locate information at specific information desks.

An iterative approach was employed during development of the prototype services. The dialogue design process consisted of two stages leading to the design of the prototypes that were evaluated during the user testing conducted within this study. The first stage involved mapping metaphors onto the service structure and functionality by using flow diagrams, which were then evaluated in a ‘quick and dirty’ style by three HCI experts. The second stage involved implementing a section of the service, which was operated using the WOZ technique. Relevant features, structural elements, and meaningful vocabulary for the metaphor-based dialogue, were all based on the POPITS data derived from preliminary study one.

The dialogues for each service were designed using relevant human factors principles for the design of phone-based interfaces. A welcome message was provided so the user was able to identify which service they had reached (Schumacher, 1992); a title was provided at the beginning of each service message to inform users of the result of their last command and of their current position within the service (Hallstead-Nussloch, 1989); menu options were presented in the ‘goal-action’ sequence, as this is more consistent with the cognitive make up of the task (Hallstead-Nussloch, 1989); the users were told how many menu options to expect in the service prompts, as a way of reducing cognitive demand and decreasing the error rate (Schumacher, 1992); the service messages were interruptible to allow more experienced users to progress more quickly through the service, reducing user frustration and increasing efficiency (Marics and Engelbeck, 1997); time-outs were implemented which came into effect if the user did not provide a response within a period of 5 seconds after the end of the service prompt. The time-outs repeated the service prompt to the user, and after three

time-outs the user was connected to the human operator (Marics and Engelbeck, 1997); and finally, control options were implemented, as recommended by Hallstead-Nussloch (1989), allowing users to navigate through the service. The following control options were implemented, and were available at all service levels, and within every dialogue prompt that was accessed:

- Repeat: repeats the service dialogue.
- Back: takes the user to the previous level of the service.
- Return: takes the user back to level 1 of the service (entry point).
- Exit: terminates the users interaction with the service.
- Help: provides the user with a full explanation of the control functions.

For the services implemented for this study, different characters (voices) were used within each service to reflect the people the user would encounter as they progressed through the corresponding real world environment. This reflected the POPITS data generated by participants, which suggested that characters playing different roles were a strong feature in their descriptions of the real world metaphor referents. These characters could also serve to orient users within each level of the service, by acting as landmarks or navigation cues. Different synthetic voices, both male and female, were used for the dialogue at different levels of the service. The complexity of the services was consistent, and both the number of levels within services, and the ordering of menu options within the service prompts was the same for all four services.

A summary of the structure, menu options at each level, and characters used to implement each of the four services can be seen in Tables 4.7 to 4.10 below. Examples of service messages and prompts from the services can be seen in Table 4.11.

Table 4.7. Features of the standard service used for preliminary study two

Features of the system	Standard service				
	Level 1	Level 2	Level 3	Level 4	Level 5
Structure	Welcome message with 2 menu options	2 menu options	3 menu options	3 menu options	No menu options
Options	Numbers 1, or 2	Numbers 1 or 2	Numbers 1, 2, or 3	Numbers 1, 2 or 3	City information
Character	A single anonymous female synthesised voice				

Table 4.8. Features of the travel system service used for preliminary study two

Features of the system	Travel system service				
	Level 1	Level 2	Level 3	Level 4	Level 5
Structure	In a car in a car park with 2 lanes leading from the car park	2 roads	3 turnings	3 signs	No menu options
Options	Left lane, and right lane	A road, or B road	First, second, or third turning on the left	Red, Blue, or Yellow sign	City information
Character	In-car navigation system	In-car navigation system	In-car navigation system	In-car navigation system	John at the tourist information centre

Table 4.9. Features of the office filing system service used for preliminary study two

Features of the system	Office filing system service				
	Level 1	Level 2	Level 3	Level 4	Level 5
Structure	In an office facing 2 filing cabinets	2 drawers	3 sections	3 folders	No menu options
Options	Left, or right filing cabinet	Bottom, or top drawer	First, second, or third section	Red, Blue, or Yellow folder	City information
Character	Kate the office manager	Kate the office manager	Kate the office manager	Kate the office manager	Folder information service

Table 4.10. Features of the shopping system service used for preliminary study two

Features of the system	Shopping system service				
	Level 1	Level 2	Level 3	Level 4	Level 5
Structure	In an elevator on the ground floor with 2 floors to choose from	2 departments	3 sections	3 information desks	No menu options
Options	Up 1 floor, or up 2 floors	Left, or right	Left, straight ahead, or right	Left, straight ahead, or right	City information
Character	Andy the lift operator	Andy the lift operator	Jo the store manager	Ben the shop assistant	Jackie at the information desk

Table 4.11. Example dialogue from the 4 services used for preliminary study two

Standard	Travel system	Office Filing System	Shopping
Option 2, Eating Out. There are 2 eating out options available. For restaurants select option 1, and for cafes select option 2. Say 1 or 2.	This is the Eating Out lane. There are 2 types of road available at this point. For cafe city take the 'A' road, and for restaurant city take the 'B' road. Say 'A' or 'B'.	This is the Eating Out filing cabinet. There are 2 drawers available. For cafe information select the bottom drawer, and for restaurant information select the top drawer. Say bottom or top.	This is the Eating Out floor. There are 2 departments available. The cafe department is on the left, and the restaurant department is on the right. Say left or right.
Option 1, Restaurants. There are 3 restaurant options available. For Chinese restaurants select option 1, for Indian restaurants select option 2, and for Italian restaurants select option 3. Say 1, 2, or 3.	This is the 'B' road for restaurant city. There are 3 turnings available at this point. For the Chinese restaurant district take the first turning on the left, for the Indian restaurant district take the second turning on the left, and for the Italian restaurant district take the third turning on the left. Say first left, second left, or third left.	This is the Restaurant drawer. There are 3 sections available. For Chinese restaurant information select the first section, for Indian restaurant information select the second section, and for Italian restaurant information select the third section. Say first, second, or third.	Hi I'm Mary the store manager. This is the Restaurants department. There are 3 sections available in this department. For the Chinese restaurant section turn left, for the Indian restaurant section go straight ahead, and for the Italian restaurant section turn right. Say left, straight ahead, or right.
Option 2, Indian Restaurants. There are 3 Indian restaurant options available. For budget Indian restaurants select option 1, for mid-range Indian restaurants select option 2, and for top-end Indian restaurants select option 3. Say 1, 2, or 3.	This is the second turning on the left for the Indian restaurant district. For budget Indian restaurants follow the red signs, for mid range Indian restaurants follow the blue signs, and for top end Indian restaurants follow the yellow signs. Say red, blue, or yellow.	This is the Indian restaurant section. There are 3 folders in this section. For information about budget Indian restaurants select the red folder, for information about mid-range Indian restaurants select the blue folder, and for information about top-end Indian restaurants select the yellow folder. Say red, blue, or yellow.	Hi I'm Ben the shop assistant. This is the Indian restaurant section. There are 3 information desks in this section. The information desk for budget Indian restaurants is on the left, the information desk for mid-range Indian restaurants is straight ahead, and the information desk for top-end Indian restaurants is on the right. Say left, straight ahead, or right.
Option 1, the budget Indian restaurants are...	Hi, I'm John. You are at the tourist information centre. The budget Indian restaurants are...	This is the red folder information service. The budget Indian restaurants are...	Hi I'm Jackie and I work on the information desk. The budget Indian restaurants are...

4.3.3 Methodology

4.3.3.1 Design

The design was a 3 condition between subjects design, and the between subjects factor was the version of the service. There were four different versions of the service, three of which were metaphor-based (shopping, office filing system, and travel system), and one of which was designed without reference to any metaphor. The non-metaphor service was designed by simply pairing numbers with menu options, and will be referred to as the standard service.

4.3.3.2 Participants

Twenty participants took part in the study, consisting of 1 male and 19 females with ages ranging from 20 to 26 (mean age = 23). The participants were divided into the three experimental groups with six participants assigned to each of the shopping and office filing system services, and eight participants assigned to the travel system service. For the control trial (trial 1), all three groups completed tasks using the standard service, and for trial 2 each group used one of the metaphor-based services.

4.3.3.3 Apparatus

A Wizard of Oz (WOZ) methodology (Fraser and Gilbert 1991) was used for the experiment. The technique involved the experimenter simulating the functionality of a fully implemented system, to create the illusion that the user was interacting with a real telephone service. Each of the services was designed as a single webpage, with each hyperlink of the webpage linked to a corresponding sound file. The experimenter clicked on hyperlinks, which played the appropriate sound files to the participants. The sound files were played through speakers connected to the computer, which was then relayed via a speakerphone to the participant at the other end of the phone line. The experimenter wore a telephone headset connected to the speakerphone, which made it easier to hear participants' responses. The computer was loaded with TrueActive monitoring software (TrueActive Corporation, Kennewick, WA) to log the sound files that the participants used, and the amount of time that was spent listening to each file.

4.3.3.4 Data collection

Subjective attitudes towards each implementation of the service were recorded after each trial using a 34-item, 7-point Likert style usability questionnaire (Appendix 9), which was balanced for both positively and negatively worded statements. The questionnaire was the predecessor of the questionnaire used for the main three experiments, and which can be seen in Appendix 2. It was evident that the 34 items were not all measuring the same construct (usability), and for the main experiments, the questionnaire was expanded to 50 items, and based on the six subjective factors proposed by Hone and Graham (2000).

Two objective measures of task performance were collected during the participants' interaction with the services: successful task completion, and the time taken to complete the tasks as a percentage of the minimum path prompt time. The second of these measures was the actual time taken to complete a task, represented as a percentage of the total prompt time if the optimum path (lowest number of nodes) had been used, but will simply be referred to as 'time' for the purposes of this study (absolute transaction times can be seen in Appendix 21).

To gather demographic data, and data about the frequency of each participants' previous mobile phone experience, previous fixed line telephone experience, and previous computing experience, a Technographic questionnaire was designed. The questionnaire consisted of 16 visual analogue response items to which participants could respond along a sliding scale ranging from 'Never' to 'Often', for example, 'How often do you use a mobile phone to access the Internet?' The scale was 100mm long, and participants responses produced interval data assigned a value between 0 and 100, with a '0' representing no experience of using the item in question, and '100' representing frequent usage. The questions were presented in three sections: mobile phone; fixed line telephone; and computer. The final questionnaire can be seen in Appendix 10, and was used for all studies and experiments conducted for this thesis. The validity of the questionnaire was established by analysing the data from 45 participants using Principal Component Analysis (PCA) to determine whether the three sections of the questionnaire yielded valid and coherent internal constructs. The results indicated that the 'computing' and 'mobile phone' components were coherent

constructs, but that the ‘fixed line telephone’ component was not. The full study that generated this data can be seen in Appendix 11.

4.3.3.5 Procedure

Each participant was tested individually within a 1-hour session. Firstly, participants completed the technographic questionnaire. Secondly, participants were called on the landline telephone in their experimental cubicle, and had to complete a practice task followed by three tasks using the standard service. Each task required participants to find a specific piece of information, for example, ‘Find the names of 2 wine bars that close at 11pm and then exit the service’. A full task list can be seen in Appendix 12. Participants were then asked to complete a usability questionnaire. Task times were recorded using a stopwatch. Finally, participants were called a second time to complete three tasks using one of the metaphor-based services. Again, task times were recorded, and on completion of the tasks, participants were asked to complete a second usability questionnaire.

4.3.4 Results

One-Way ANOVA tests were used to compare differences in age, previous mobile phone experience, previous fixed line telephone experience, and previous computing experience between the three experimental groups. No significant differences were found between groups, suggesting that participant groupings did not vary on these factors prior to the study. Table 4.12 shows descriptive data for participants’ age, and previous telephone and computing experience. For telephone and computing experience measures, a higher score indicated greater experience levels.

Table 4.12. Descriptive data for the individual difference measures

Participant variable	Minimum	Maximum	Mean	Standard deviation
Age	20.00	32.00	21.55	3.19
Mobile phone experience	8.14	58.86	33.77	12.34
Telephone experience	2.40	56.80	23.51	14.86
Computing experience	17.50	91.50	66.10	17.70

4.3.4.1 Subjective measures

Three paired samples t-tests were performed to investigate differences in subjective measures of attitude between the standard service and each of the three metaphor-

based services. The only significant difference was between the travel system service and the standard service (Table 4.13), suggesting that participants perceived the usability of the travel system service to be significantly lower. To investigate differences in usability between the three metaphor-based services, a one-way ANOVA was performed (across services and trials: mean attitude score = 3.91, SD = 0.69). Although no significant differences were found, the office filing system service scored the highest usability mean, which was only 2% lower than the mean score for the standard service, compared to a 4 % lower score for the shopping service group, and a 20% lower score for the travel system service group.

Table 4.13. Paired t-test results for attitude between the metaphor and non-metaphor services

Experimental group	Trial number	Mean	SD	t	df	Sig. (2-tailed)
Shopping service	1 (standard)	3.57	0.84	0.42	5	n.s.
	2	3.42	0.82			
Office filing system	1 (standard)	4.42	0.77	0.19	5	n.s.
	2	4.33	0.92			
Travel system	1 (standard)	4.30	0.83	2.38	7	<0.05
	2	3.44	0.74			

4.3.4.2 Objective measures

Differences in ‘time’ between the standard service and the three metaphor-based services were explored using paired samples t-tests. Although no significant differences were found, the time taken to complete tasks was lower for all of the metaphor-based services than it was when using the standard service (see Table 4.14).

Table 4.14. Paired t-test results for ‘time’ between the metaphor and non-metaphor services

Experimental group	Trial number	Mean	SD	t	df	Sig. (2-tailed)
Shopping service	1 (standard)	92.39	6.14	2.12	5	n.s.
	2	75.44	11.36			
Office filing system	1 (standard)	105.00	39.12	2.29	5	n.s.
	2	82.92	23.77			
Travel system	1 (standard)	94.56	28.54	0.56	7	n.s.
	2	89.05	32.47			

In order to investigate whether there was a significant difference in ‘time’ between the three metaphor-based services, a one-way ANOVA was performed (across services

and trials: mean time = 90.08, SD = 23.50). No significant differences were apparent ($F(2,17)=0.50$; $p=0.62$), but the greatest improvement in performance was shown by the office filing system service, with participants taking 22% less time to complete tasks than when they used the standard service, compared to 17% less time for the shopping service, and 6% less time for the transport system service. Successful task completion rates were high across all three groups because participants tended to persevere until they found the relevant information. However, participants using the office filing system service were the only group to achieve 100% successful task completion across all three tasks.

4.3.5 Discussion

Of the metaphors examined, the office filing system service emerged as being the service that was perceived most positively, and generated the best performance levels, whereas the travel system service was perceived most negatively. However, none of the metaphor services were rated as being significantly better than a standard phone service. Since contemporary number-based services (such as the standard service) are so well established, it may be necessary for users to invest time and effort to learn and accept new metaphor-based services. It may therefore be considered encouraging that the usability of the office filing system service was rated as being similar to the standard service, given the novel nature of this system and the limited exposure afforded to participants in the present study. It was therefore decided that, in experiment one, the usability of the office filing system service would need to be examined over longer periods of time, if performance benefits were to become evident.

With reference to the final objective of the study, participant feedback suggested that some general changes needed to be made to the design of the prototypes. Firstly, only one voice should be used, as the different voices confused, rather than aided, navigation, and were also perceived as being of different qualities, which may have affected perceptions of the services. When asked which voice was preferred, participants unanimously opted for one of the female voices, which was subsequently used to implement the prototypes for future experiments. Secondly, participants found the synthesized speech difficult to comprehend at the start, and it was suggested that the addition of sound effects would simply compound this problem. It was therefore

decided not to integrate sound effects into future versions of the services. Thirdly, participants preferred the services that had different menu options at each level, as it helped them to differentiate between levels, and subsequently to orient themselves within the service structure. It was therefore decided to make the menu options for the metaphor-based services different at each level of the service. Finally, participants suggested that a global command be implemented to return them to the start of the service from any point within the service, and for this to be given a meaningful name. Such a function was provided in the prototypes, but was different for each service, and confused participants. For example, for the travel service, the command was ‘car park’, for the office filing system service it was ‘office’, and for the shopping service it was ‘ground floor’. In order to be less ambiguous, the function was named ‘return’ in successive prototypes. Specific feedback was also provided for individual services, such as participant’s disliking for the coloured signposts at level four of the travel system service, and for the information desks at level four of the shopping service. The general and specific feedback was factored into the redesigned prototypes used in experiment one.

4.4 Chapter summary

Chapter four described the two preliminary studies that utilised a HCD process to generate, select, and develop interface metaphors, which were used to implement different versions of a speech-based automated mobile city guide service. The services were then evaluated in a user testing study. The main findings of the chapter were that:

- Card-sorting provides a usable and productive methodology for selecting and developing interface metaphors
- The POPITS model for GUI metaphor design is an effective framework for developing metaphors for speech-based systems, and that when used in conjunction with a card-sorting methodology, produces more relevant metaphor features
- Generally, participants were able to use metaphor to spontaneously explain how to perform tasks

- Metaphors generated using card-sorting and sketching may be applicable to a range of service domains
- Five categories of metaphor may be most appropriate to the design of speech-based mobile phone services
- The office filing system metaphor service may lead to improved levels of attitude and performance compared both to other metaphor-based services, and to non-metaphor services.

:: CHAPTER 5

Spatial metaphors for a speech-based mobile city guide service

5.1 Introduction

The first main experiment reported in this chapter of the thesis focused on full prototype versions of the speech-based mobile city guide service revised as a result of the feedback from preliminary study two. The current experiment differs from the Dutton et al. (1999) experiment in that three metaphors were used; the metaphors were generated as part of a HCD process; a mobile city guide service was used rather than a home shopping service; the service was more extensive, consisting of a deeper hierarchy, and enabling the user to access more information; and finally the service was designed to be accessed from a mobile phone using speech input, rather than a fixed line telephone using keypad input.

The two main objectives of this experiment were:

1. To investigate whether different metaphor-based versions of an automated mobile city guide service led to improvements in usability compared to a version of the service that was not based on any underlying metaphor.

2. To investigate whether there were any differences in usability between different metaphor-based versions of an automated mobile city guide service.

It was expected that the design of services based on spatial interface metaphors would lead to improved perceptions towards and performance with those services compared to a non-metaphor version of the service.

5.2 Prototype design and development

The service was an extended version of the service used in preliminary study two, with one menu item added to level two of the hierarchy. This resulted in an additional 18 information nodes at level five of the service, which consisted of a total of 54 information nodes.

As with preliminary study two, one version of the service was designed in the same style as most commercially available automated telephone services, for example ‘For option 1, say 1, for option 2, say 2...’, and was referred to as the standard service. The designs of the other three versions of the service were based on the 3 different metaphors used in preliminary study two. These metaphors were a travel system, in which participants were required to navigate through different districts and zones of a city to find an appropriate parking space; an office filing system, in which participants were required to select from different drawers, partitions and coloured folders in a row of filing cabinets; and a shopping metaphor, in which participants could move around different floors and aisles of a department store to locate information.

Before the prototypes were used for this experiment, a pilot study was conducted which involved assigning one participant to each service type, and testing the services in realistic mobile contexts. This ecologically valid testing of the prototypes produced the following outcomes:

1. In mobile environments, for example a shopping mall, the speed and efficiency of the services was cited by participants as being an important factor. Therefore, in order for the metaphor-based services to be efficient without seeming abrupt or rude, a limit of approximately 30 seconds was imposed on dialogue length.

2. The omission of sound effects was confirmed as a good design decision because participants in noisy mobile environments were sometimes struggling just to hear and to understand the service voice, and any additional sounds within the service may have been distracting and possibly confusing.

The ordering of menu options within the service prompts was the same for all services. The 3 constituent parts of each dialogue were also balanced across the services as percentages of the total dialogue time:

1. The service message, which is the part of the dialogue that informs the user of their current location within the service structure.
2. The service prompt, which is the part of the dialogue where the user is given menu options to choose from in order to progress through the service.
3. The control options.

There were no changes made to the standard service, but the changes made to the metaphor-based prototypes as a result of participant feedback from preliminary study two will now be outlined.

The changes made to the travel system service:

- Previously, when accessing the service at level one, the participant was in a car park, and had to choose a left, or a right lane leading from the car park. For the current prototype the participant is on a three-lane road system, and has to choose either the middle, inside, or outside lane.
- Previously, at level two, the participant had to choose from either an 'A' road, or a 'B' road. For the current prototype the participant has to select from different city zones by taking either the first, or the second turning on the left.
- Previously, at level three, the participant had to choose from either the first, second, or third turning on the left. For the current prototype the participant has to select from different roads to park on: Station Road, Seaside Road, or Central Road.

- Previously, at level four, the participant had to choose from either a red, blue, or yellow sign. For the current prototype the participant has to select from different parking meters: first, second, or third.
- Previously, at level five, 'John' at the tourist information centre provided the city information to the participant. For the current prototype the participant is told the information by the automated car parking meter.

The changes made to the office filing system service:

- Previously, at level three, the participant had to choose from either the first, second, or third section of a filing cabinet drawer. For the current prototype the participant has to select from different partitions: front, middle, or rear.

The changes made to the shopping system service:

- Previously, when accessing the service at level one, the participant had to choose to go either 'up one floor', or 'up two floors' in the elevator. For the current prototype the participant has to choose to go to: ground floor, first floor, or the basement.
- Previously, at level two, the participant had to choose from either the department on the left, or the right, on leaving the elevator. For the current prototype the participant has to select from different sections: left, or right.
- Previously, at level three, the participant had to choose from either the section on the left, on the right, or straight ahead. For the current prototype the participant has to select from different aisles: first, second, or third.
- Previously, at level four, the participant had to choose from either the information desk on the left, on the right, or straight ahead. For the current prototype the participant has to select from different shelves: top, middle, or bottom.
- Previously, at level five, 'Jackie' at the information desk provided the city information to the participant. For the current prototype the participant is told the information by the shelf stacker.

A summary of the structure, menu options at each level, and characters (voices) used to implement each of the four services can be seen in Tables 5.1 to 5.4 below.

Table 5.1. Features of the standard service used for experiment one

Features of the service	Standard Service				
	Level 1	Level 2	Level 3	Level 4	Level 5
Structure	Welcome message with 3 menu options	2 menu options	3 menu options	3 menu options	No menu options
Options	Numbers 1, 2, or 3	Numbers 1 or 2	Numbers 1, 2, or 3	Numbers 1, 2 or 3	City information
Character	Anonymous	Anonymous	Anonymous	Anonymous	Anonymous

Table 5.2. Features of the travel system service used for experiment one

Features of the service	Travel System Service				
	Level 1	Level 2	Level 3	Level 4	Level 5
Structure	In a car on a 3 lane road system driving through the city	2 zones	3 different roads	3 parking meters	No menu options
Options	Middle, inside, or outside lane	First or second turning on the left	Station road, Seaside road, or Central road	First, second, or third parking meter	City information
Character	In-car navigation system	In-car navigation system	In-car navigation system	In-car navigation system	Automated car parking meter

Table 5.3. Features of the office filing system service used for experiment one

Features of the service	Office Filing System Service				
	Level 1	Level 2	Level 3	Level 4	Level 5
Structure	Standing in the main office facing 3 filing cabinets	2 drawers	3 partitions	3 folders	No menu options
Options	Left, middle, or right filing cabinet	Bottom or top drawer	Front, middle, or rear partition	Red, blue, or yellow folders	City information
Character	The office manager	The office manager	The office manager	The office manager	The folder information service

Table 5.4. Features of the shopping service used for experiment one

Features of the service	Shopping Service				
	Level 1	Level 2	Level 3	Level 4	Level 5
Structure	Standing in the elevator with 3 floors to choose from	2 sections	3 aisles	3 shelves	No menu options
Options	Ground floor, first floor, or basement	Section to the left or right on leaving elevator	First, second, or third aisle	Top, middle, or bottom	City information
Character	The elevator operator	The elevator operator	The store manager	The customer advisor	The shelf stacker

5.3 Methodology

5.3.1 Design

The experimental design was a 4x4 mixed factorial design with 2 independent variables. The between subjects factor was the type of service, and the within subjects factor was the trial.

Each experimental group completed a control trial, and then 3 experimental trials, with trials being held on 4 consecutive days. Participants attempted 3 tasks during each trial. For the control trial, all experimental groups completed tasks using the standard service. For trials 1 to 3, experimental group A continued to use the standard service, experimental group B used the travel system metaphor service, experimental group C used the office filing system metaphor service, and experimental group D used the shopping metaphor service. Two types of dependent variable were evaluated: (1) performance with the services, and (2) attitude towards the services.

5.3.2 Participants

A total of sixty participants was recruited, consisting of 28 males and 32 females with ages ranging from 18 to 48 (mean = 25). The participants were divided into the 4 experimental groups balanced for age and gender, with 15 participants in each group. Participants were drawn from a range of educational (graduate and non-graduate) and vocational (professional and non-professional) backgrounds.

5.3.3 Apparatus

A Wizard of Oz (WOZ) methodology (Fraser and Gilbert 1991) was used for the experiment. As in preliminary study two of this thesis, each of the services was designed as a single webpage. Each hyperlink of the webpage linked to a corresponding sound file, which, when clicked, played the appropriate sound files to the participants. The computer was loaded with TrueActive monitoring software (TrueActive Corporation, Kennewick, WA) to log the sound files that the participants used.

5.3.4 Data collection

5.3.4.1 Subjective measures

Subjective attitudes towards each implementation of the service were gathered using a Likert questionnaire (see Appendix 2), which was a fully revised and extended version of the questionnaire used in preliminary study two of this thesis, and was discussed in chapter three. The questionnaire was administered after each of the four experimental testing sessions. Participants responded to the 50 usability statements, with a 7-point response scale ranging from ‘strongly agree’ to ‘strongly disagree’. Scores were recoded so that a high score for a factor indicated a positive attitude, and a low score a more negative attitude. Table 5.5 provides a description of each factor.

Table 5.5. Attitude measures for experiment one

Measure	Description
System response accuracy	Refers to how accurately the service recognised user input, and provided a response that matched the user’s expectations
Likeability	Refers to whether the user found the service useful and pleasant and whether they would choose to use it again
Cognitive demand	Refers to the users opinion about the amount of effort necessary to use the service, and how they felt as a result of expending this effort
Annoyance	Refers to the degree to which the user found the service repetitive, boring, irritating or frustrating
Habitability	Refers to whether the user’s conceptual model of the service was sufficient to inform them of what to do and what the service was doing
Speed	Refers to how quickly the user perceived the service as responding to their input, and to the overall duration of the interaction

5.3.4.2 Objective measures

Performance data were derived from the monitoring software log files. A total of 8 objective measures of task performance were collected during participant interactions with each service (absolute transaction times were not analysed, but can be seen in Appendix 21). Table 5.6 provides a description of each measure.

Table 5.6. Performance measures for experiment one

Measure	Description
Time	The time to complete a task was calculated as a percentage of the total length of the service dialogue (a lower score was taken to indicate a better performance)
Task completion	The number of times the tasks were successfully completed as a percentage of the total tasks (a higher score was taken to indicate a better performance)
Prompt interrupts	The number of times the dialogue was interrupted with a response was calculated as a percentage of the total number of dialogue nodes used (a higher score was taken to indicate a better performance)
Total nodes	The number of dialogue nodes used to complete a task (a lower score was taken to indicate a better performance)
Optimum nodes	The number of dialogue nodes used to complete a task was calculated as a percentage of the optimum number – the shortest route (a lower score was taken to indicate a better performance)
No user response	Failure to respond to a service prompt after a period of 5 seconds (a lower score was taken to indicate a better performance)
Returns	The number of times a request to return to the first level of the service was made (a lower score was taken to indicate a better performance)
Repeats	The number of times a request to repeat a dialogue prompt was made (a lower score was taken to indicate a better performance)

5.3.5 Procedure

The experiment was conducted in two parts. The first part was a preliminary session during which participants completed the technographic questionnaire (see Appendix 10), which was discussed in section 4.3.3.4 of chapter four. They also provided contact details, and time availability information, and were given the experimental materials needed to complete the experiment: a participant information form outlining the experimental procedure; a response scale to refer to when responding to the Likert questionnaire statements; a practice task sheet, and a task sheet containing the 3 tasks to be completed during each trial (Appendix 13). Each task required the participants to find a specific piece of information. In order to reduce the potential for ceiling effects the tasks became progressively harder within each testing session. This ensured that the best performing participants found the hardest tasks moderately difficult. Examples of typical tasks can be seen below.

- Find the names of 2 wine bars that close at 11pm and then exit the service (easiest task).
- Find the names of 2 mid-range Italian restaurants, and then find the names of 2 comedy films showing in the evenings, and then exit the service (hardest task).

The sequence of correct menu options required to successfully complete tasks was balanced across the 4 experimental groups so that differences between groups for task completion times and prompt interrupt rates (objective measures 1 and 3 respectively) could not be attributed to the ordering of menu options. In addition, the ordering of menu options to successfully complete tasks was randomised across experimental trials. Participants were informed that:

- All of the information needed to use the services would be provided within each service, including a separate ‘Help’ function
- They would be given a practice task before the experiment started
- All responses to the service prompts had to be spoken rather than keypad activated
- They would be allowed to interrupt the service prompts with a response at any time.

For the second part of the experiment, participants were called on their mobile phones and had to complete tasks using the city guide service, but before attempting the control trial tasks, they were given a practice task. The experimenter activated the service messages and prompts, and the participants responded with their menu selections. On completion of all 3 experimental tasks, participants were asked to verbally respond to the Likert questionnaire statements by referring to the 7-point response scale. The Likert questionnaire was administered in this way to guarantee that all responses were given immediately after using the service. Finally, participants were interviewed for 5 minutes regarding their experience of using the service, and on successive trials participants using the metaphor-based services were asked to compare the service with the standard service they used for the first trial. The questions used for this semi-structured interviewing can be seen in Appendix 14. This procedure was repeated over the following three consecutive days of the experiment.

5.4 Results

A total of 240 phone calls were made across the 4 experimental trials, and participants answered these calls from a range of 16 different locations. Table 5.7 shows the full range of locations, and the percentage of calls that were answered in each.

Table 5.7. Mobile phone call locations for experiment one

Location of the call	Frequency	Percentage
Bedroom	96	40.0
Lounge	58	24.2
Office	22	9.2
Kitchen	14	5.8
Common Room	9	3.8
Corridor	9	3.8
Car Park	5	2.1
Train	5	2.1
Café	4	1.7
Bar	4	1.7
Park	4	1.7
Computer Lab	3	1.3
Campus – outside	3	1.3
Library	2	0.8
Supermarket	1	0.4
Bus	1	0.4

One-Way ANOVA tests were used to compare differences in age, previous mobile phone experience, previous fixed line telephone experience, and previous computing experience between the four experimental groups. No significant differences were found between groups, suggesting that participant groupings did not vary on these factors prior to the study. Table 5.8 shows descriptive data for participants' age, and previous telephone and computing experience.

Table 5.8. Descriptive data for the individual difference measures for experiment one

Participant variable	Minimum	Maximum	Mean	Standard deviation
Age	18.00	48.00	25.18	7.48
Mobile phone experience	6.71	58.71	30.69	12.95
Telephone experience	0.00	71.60	25.78	17.35
Computing experience	3.50	100.00	69.34	23.45

Table 5.9 shows the range of possible scores for the performance and attitude measures, as well as the grand mean and standard deviation for all users and across all services.

Table 5.9. Score range and means for the performance and attitude measures for experiment one

Measure type	Measure	Range	Mean	SD
System performance measures	Time	0-100	72.61	7.89
	Task completion	0-100	94.56	7.41
	Prompt interrupts	0-100	75.79	7.79
	Total nodes	>27	31.13	2.02
	Optimum nodes	>100	113.97	7.44
	No user response	>0	0.09	0.23
	Returns	>0	0.98	0.69
	Repeats	>0	0.53	0.63
Attitude measures	System response accuracy	1-7	5.13	0.77
	Likeability	1-7	4.31	0.76
	Cognitive demand	1-7	4.94	0.85
	Annoyance	1-7	3.63	0.86
	Habitability	1-7	4.67	0.91
	Speed	1-7	3.96	0.90

5.4.1 Subjective attitude measures

A 3x4 mixed design MANCOVA (trial x experimental group) was used to explore differences between experimental groups, and between experimental trials (testing days 2-4) for the 6 subjective measures of system performance. For this analysis subjective evaluations for each measure on the control trial (testing day 1) were used as a covariate, and the trials refer to the following testing days: trial 1 = day 2, trial 2 = day 3, and trial 3 = day 4. The multivariate F value for the interaction between trial and experimental group was significant ($F(18, 127.77)=1.84, p=0.03, \text{Wilks' } \lambda = 0.52$), suggesting that changes in attitude over time did not occur at the same rate for all experimental groups. The univariate results for individual subjective factors can be seen in Table 5.10 below.

Table 5.10. Main effects for subjective measures (Experimental group: A = standard service; B = travel system service; C = office filing system service; D = shopping service)

Subjective measure	Experimental group	Mean and SD	Trial 1	Trial 2	Trial 3	Main effect between trials	
						F	Sig.
System response accuracy	A	Mean	5.16	5.12	5.15	0.04	n.s.
		SD	1.24	1.10	1.09		
	B	Mean	4.58	4.87	5.19	4.62	<0.01
		SD	0.73	0.92	0.82		
	C	Mean	4.90	5.19	5.45	3.32	<0.05
		SD	0.80	0.56	0.66		
	D	Mean	5.18	5.36	5.40	0.65	n.s.
		SD	0.84	0.74	0.63		
Likeability	A	Mean	4.18	4.30	4.29	0.47	n.s.
		SD	0.97	0.99	1.00		
	B	Mean	3.85	4.04	4.26	5.40	<0.01
		SD	0.71	0.75	0.64		
	C	Mean	4.25	4.76	4.75	9.39	<0.01
		SD	0.76	0.59	0.65		
	D	Mean	4.26	4.31	4.52	3.34	<0.05
		SD	0.81	0.81	0.74		
Cognitive demand	A	Mean	5.00	5.11	5.17	0.35	n.s.
		SD	1.11	1.18	1.02		
	B	Mean	4.33	4.64	5.02	5.96	<0.01
		SD	0.71	0.94	0.87		
	C	Mean	4.41	4.99	5.31	9.11	<0.01
		SD	1.24	0.89	0.79		
	D	Mean	4.69	5.22	5.40	6.01	<0.01
		SD	0.94	0.73	0.62		
Annoyance	A	Mean	3.57	3.33	3.56	1.50	n.s.
		SD	0.98	1.04	1.09		
	B	Mean	3.61	3.50	3.81	2.67	n.s.
		SD	0.67	0.95	0.70		
	C	Mean	3.54	3.72	3.86	1.00	n.s.
		SD	0.91	0.91	1.10		
	D	Mean	3.53	3.56	3.97	4.79	<0.01
		SD	1.22	1.06	1.07		
Habitability	A	Mean	4.58	4.50	4.63	0.04	n.s.
		SD	1.36	1.10	0.88		
	B	Mean	4.29	4.56	4.67	1.07	n.s.
		SD	0.90	0.76	1.13		
	C	Mean	4.32	5.15	5.14	5.14	<0.01
		SD	0.90	2.51	0.81		
	D	Mean	4.42	4.78	4.99	2.27	n.s.
		SD	0.99	0.69	0.61		
Speed	A	Mean	3.83	3.85	3.98	0.39	n.s.
		SD	0.95	1.18	1.21		
	B	Mean	3.88	3.85	3.90	0.12	n.s.
		SD	0.74	1.03	0.97		
	C	Mean	4.38	4.30	4.12	0.91	n.s.
		SD	1.01	1.11	1.24		
	D	Mean	4.08	3.58	3.77	2.05	n.s.
		SD	0.80	0.87	0.88		

5.4.1.1 Differences between groups

In order to test whether participants had a more positive attitude towards the metaphor-based services than the standard service, the main effects of experimental group were examined for experimental trials 1 and 3 (days 2 and 4), which corresponded to their first use and last use of the services. This enabled participants' initial reactions to be recorded, and then their attitudes after using the services on 3 consecutive days.

For trial 1 there was a significant main effect between groups for cognitive demand ($F(3)=3.39$, $p=0.02$), but no significant effects for the other five subjective measures. Bonferroni pairwise comparisons showed a significant difference between the standard service and both the travel system service ($p=0.01$) and the office filing system service ($p=0.01$). These differences can be seen in Figure 5.1, and show that participants were more positive about the standard service for trial 1. Whilst no other statistically significant differences were found on trial 1, it was noted that the standard service had the highest mean scores for only 2 of the 6 subjective measures (cognitive demand, habitability). The shopping service had the highest mean scores for 2 subjective measures (system response accuracy, likeability). The travel system service had the highest mean score for 'annoyance', and the office filing system service had the highest mean score for 'speed'.

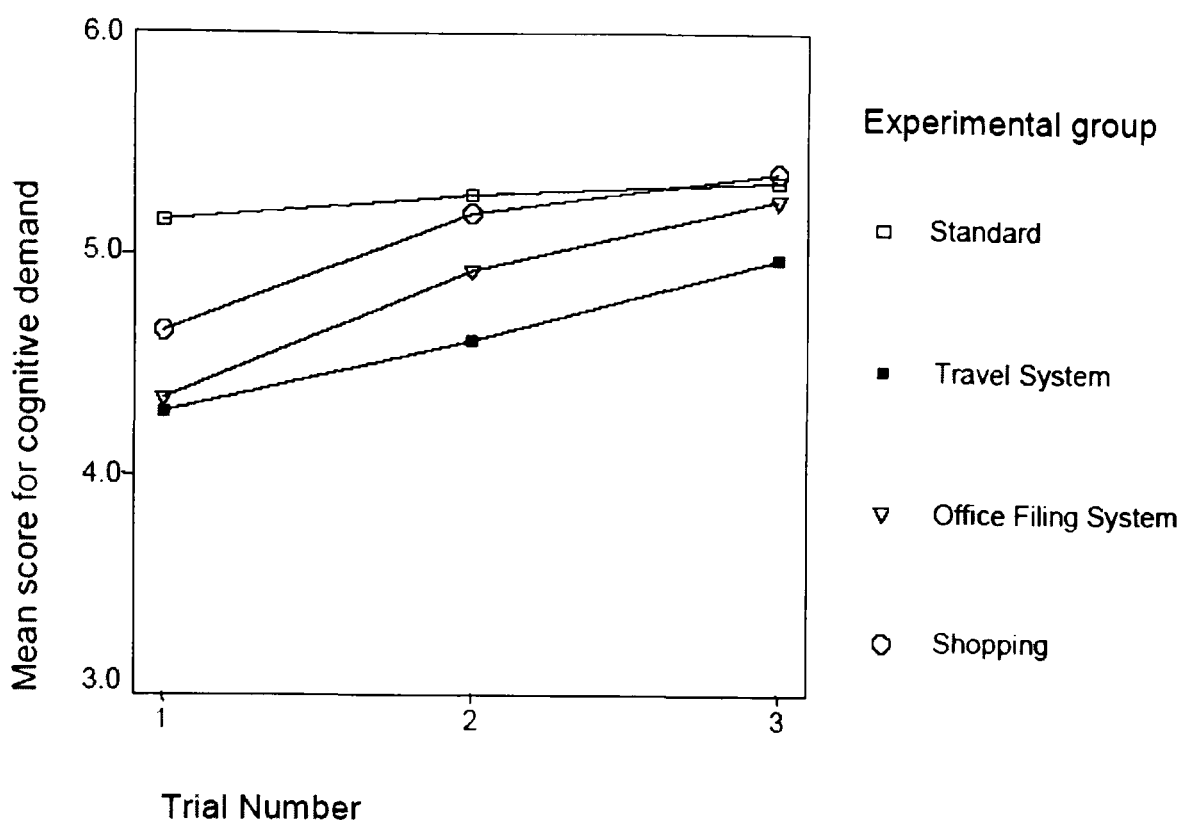


Figure 5.1. Mean scores across experimental trials for: Cognitive Demand

These results suggest that on first exposure, participants perceived the standard service as being significantly less cognitively demanding than both the road system service and the office filing system service. Overall, the metaphor-based services scored 4 of the 6 highest subjective means, two of which were generated by the shopping service (Table 5.10).

For trial 3, there were no significant main effects between groups for any of the 6 subjective measures, which shows that after 3 exposures the metaphor-based services were perceived as positively as the standard service. The office filing system service had the highest means for 4 of the 6 subjective measures: system response accuracy, likeability, habitability, and speed. The shopping service had the highest means for the other 2 subjective measures: cognitive demand, and annoyance. The standard service recorded the lowest means for 3 of the 6 measures: system response accuracy, annoyance, and habitability (Table 5.10).

These results suggest that after using the services on 3 consecutive days, participants perceived the office filing system service most positively, and the standard service least positively.

5.4.1.2 Differences between trials

In order to determine whether participants' attitudes towards the metaphor-based services improved more over time than they did with the standard service, the main effects of trial were examined for each experimental group (Table 5.10). For the standard service, there were no significant improvements between trials for any of the subjective measures. For the travel system service, there was a significant improvement between trials for 3 subjective measures: system response accuracy, likeability, and cognitive demand. For the office filing system service, there was a significant improvement between trials for 4 subjective measures: system response accuracy, likeability, cognitive demand, and habitability. For the shopping service, there was a significant improvement between trials for 3 subjective measures: likeability, cognitive demand, and annoyance. The office filing system service therefore showed significant improvements over time for the highest number of subjective measures.

5.4.2 Objective performance measures

A 3x4 mixed design MANCOVA (trial x experimental group) with the control trial score as a covariate for each measure was used to explore differences between experimental groups, and between experimental trials for the 8 objective measures. The multivariate F value for the interaction was not significant ($F(24, 52.81)=0.77$, $p=0.75$, Wilks' $\lambda = 0.42$), suggesting that changes in performance over time occurred at the same rate for all services. The univariate results for individual performance measures can be seen in Table 5.11 below.

Table 5.11. Main effects for objective measures (Experimental group: A = standard service; B = travel system service; C = office filing system service; D = shopping service)

Objective measure	Experimental group	Mean and SD	Trial 1	Trial 2	Trial 3	Main effect between trials	
						F	Sig
Time as a percentage of prompt time	A	Mean	71.96	77.57	69.40	13.44	<0.01
		SD	8.59	8.23	6.94		
	B	Mean	76.78	73.34	63.78	21.72	<0.01
		SD	10.22	10.08	9.19		
	C	Mean	81.41	74.71	67.33	19.01	<0.01
		SD	16.08	9.76	8.75		
	D	Mean	73.88	73.69	63.69	17.43	<0.01
		SD	5.02	4.40	5.03		
Percentage correct task completion	A	Mean	97.33	98.67	97.33	0.20	n.s.
		SD	7.04	5.16	7.04		
	B	Mean	89.33	92.00	98.67	4.40	<0.05
		SD	26.04	12.65	5.16		
	C	Mean	90.67	94.67	100.00	3.16	<0.05
		SD	18.31	9.15	0.00		
	D	Mean	82.67	96.00	97.33	3.61	<0.05
		SD	19.81	8.28	7.04		
Percentage prompt interrupts of the total nodes	A	Mean	76.96	80.03	80.45	1.24	n.s.
		SD	7.97	8.50	5.73		
	B	Mean	70.59	78.78	80.10	7.58	<0.01
		SD	10.05	6.86	7.36		
	C	Mean	65.35	75.81	78.22	11.91	<0.01
		SD	17.60	11.04	8.69		
	D	Mean	71.19	74.90	78.57	3.28	<0.05
		SD	7.32	10.23	6.79		
Number of 'no user responses'	A	Mean	0.33	0.00	0.07	6.39	<0.01
		SD	0.49	0.00	0.26		
	B	Mean	0.14	0.00	0.00	0.67	n.s.
		SD	0.53	0.00	0.00		
	C	Mean	0.31	0.08	0.00	0.85	n.s.
		SD	1.11	0.28	0.00		
	D	Mean	0.13	0.07	0.00	1.58	n.s.
		SD	0.35	0.26	0.00		
Total number of nodes	A	Mean	31.00	29.50	29.67	0.50	n.s.
		SD	4.86	1.38	2.88		
	B	Mean	31.00	31.60	31.40	0.08	n.s.
		SD	1.00	1.52	2.61		
	C	Mean	33.00	30.50	28.50	2.31	n.s.
		SD	3.69	1.87	2.07		
	D	Mean	28.25	32.00	31.00	2.34	n.s.
		SD	2.50	0.00	2.45		
Number of nodes as a percentage of the optimum number of nodes	A	Mean	114.81	105.36	109.88	1.64	n.s.
		SD	17.99	4.92	10.65		
	B	Mean	114.81	112.86	116.30	0.27	n.s.
		SD	3.70	5.42	9.66		
	C	Mean	122.22	108.93	105.56	2.72	n.s.
		SD	13.66	6.68	7.68		
	D	Mean	104.63	114.29	114.81	1.09	n.s.
		SD	9.26	0.00	9.07		

Number of 'Return' function	A	Mean	1.27	1.47	1.07	2.11	n.s.
		SD	0.88	0.74	0.70		
	B	Mean	0.71	0.86	0.79	0.19	n.s.
		SD	0.61	0.95	0.80		
	C	Mean	0.92	0.77	0.92	0.35	n.s.
		SD	0.95	0.83	0.76		
	D	Mean	0.80	1.07	0.93	0.75	n.s.
		SD	0.86	0.96	0.59		
Number of 'Repeat' function	A	Mean	0.47	0.27	0.40	0.44	n.s.
		SD	0.74	0.46	0.51		
	B	Mean	1.00	0.50	0.36	2.42	n.s.
		SD	1.71	0.65	0.84		
	C	Mean	0.62	0.31	0.00	2.19	n.s.
		SD	0.96	0.48	0.00		
	D	Mean	1.07	0.60	0.47	2.00	n.s.
		SD	1.28	0.91	0.74		

5.4.2.1 Differences between groups

In order to test whether participants' performance was better with the metaphor-based services than the standard service, the main effects of experimental group were examined for trials 1 and 3.

For trial 1 there was a significant main effect between groups for 'time as a percentage of prompt time' ($F(3)=2.89$, $p=0.04$), but no significant effects for the other seven objective measures. Bonferroni pairwise comparisons showed a significant difference between the standard service and the office filing system service ($p=0.03$), and between the office filing system service and the shopping service ($p=0.01$). These differences can be seen in Figure 5.2. In addition, on trial 1 the standard service showed the best mean performance levels for 5 of the 8 objective measures: time as a percentage of prompt time, percentage correct task completion, percentage prompt interrupts of the total nodes, number of return function, number of repeat function. The shopping service showed the best performance levels for the remaining 3 of the 8 objective measures (Table 5.11).

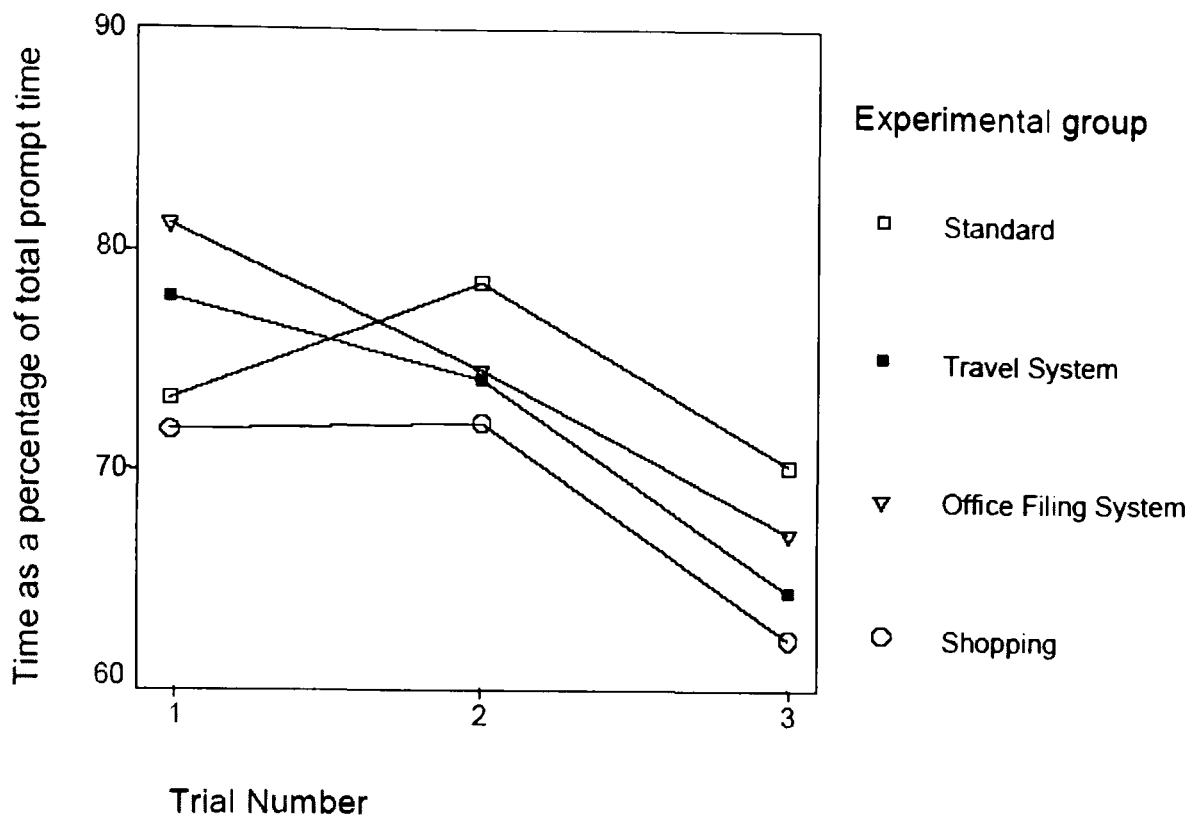


Figure 5.2. Mean scores across experimental trials for: Time as a Percentage of Total Prompt Time

The results suggest that on first exposure, participants using the office filing system service took significantly longer to perform tasks than those using both the standard service and the shopping service. Moreover, an examination of mean values suggests that performance with the standard service was best for the greatest number of measures, and that performance with the shopping service was better than performance with the other metaphor services.

For trial 3, there was a significant main effect between groups for ‘time as a percentage of prompt time’ ($F(3)=3.43, p=0.02$). Bonferroni pairwise comparisons showed a significant difference between the standard service and both the travel system service ($p=0.03$) and the shopping service ($p<0.01$). With respect to performance means, the office filing system service showed the best performance levels for the greatest number of objective measures, 5 of the total of 8 measures: percentage correct task completion, number of no user responses, total number of nodes, number of nodes as a percentage of the optimum, number of repeat function. In addition, the office filing system service was the only service to achieve a 100% correct task completion rate, and a zero use of the ‘repeat’ function. The standard service had the best performance levels for only 2 of the objective measures:

percentage prompt interrupts of the total nodes, and number of return function (Table 5.11).

The results show that after 3 exposures, participants using the standard service took significantly longer to complete tasks than those using both the travel system service and the shopping service. Although not statistically significant, an analysis of the means suggests that after 3 separate trials participants using the office filing system service performed better on more measures than those using both the standard service, and the other metaphor services.

5.4.2.2 Differences between trials

In order to determine whether participants' performance with the metaphor-based services improved more over time than it did with the standard service, the main effects of trial were examined for each experimental group (Table 5.11). For the standard service, there was a significant improvement between trials for 2 of the measures: time as a percentage of prompt time, and number of no user responses. For the travel system service, the office filing system service, and the shopping service there was a significant improvement between trials for the same 3 measures: time as a percentage of prompt time, percentage correct task completion, and percentage prompt interrupts of the total nodes.

These results show that similar improvements in performance over time were recorded for all 3 metaphor-based services. The standard service showed fewer significant improvements over time than the metaphor-based services.

5.5 Discussion

The first result to emerge from this experiment was that participants displayed no significant preferences for the metaphor-based services on the first trial. After repeated use, there were still no significant differences in attitude. However, the office filing system service scored the highest mean scores for the greatest number of attitude measures, and the standard menu style service scored the lowest mean scores for the greatest number of measures. It may therefore be concluded that participants

tended to be more positive about the office filing system service, and least positive about the standard service.

For the first trial, participants using the standard service achieved the best mean scores for the greatest number of performance measures. After repeated use, participants using the travel system service and the shopping service performed tasks significantly faster than those using the standard service. Despite this, the office filing system service achieved the best mean scores for the greatest number of performance measures. First time performance was therefore better with the standard service, but after repeated use performance was better with the office filing system service. Whilst this may suggest differences to occur with respect to performance and preference between metaphor and non-metaphor based services, differences between metaphor-based services were less apparent.

There were no significant preferences for any of the metaphor-based services on the initial trial or after repeated use. However, at the end of testing the office filing system service achieved the best mean scores for the greatest number of attitude measures, suggesting that participants were more positive about this service than the other metaphor-based services.

For first time users, the shopping service achieved the best mean scores for the greatest number of performance measures. After repeated use, there were no significant differences in performance between the metaphor-based services. However, the office filing system service achieved the best mean scores for the greatest number of performance measures, and was the only service to register optimum scores for 2 of these measures. It may therefore be argued that first time performance was better with the shopping service, but after repeated use performance was better with the office filing system service.

That attitudes towards and performance with the standard service were better than with the metaphor-based services on the first trial may not be surprising. Since participants may have been initially more familiar with the structure of the standard service, but not with the structure of the metaphor-based services, they would have to invest more time and effort learning the more novel metaphor service.

This may also explain the fact that more significant improvements in attitude and performance were observed with metaphor-based services as opposed to the standard service. As participants were already familiar with the standard service, improvements over time may be expected to be small. However, once participants had learned the metaphor-based services, the assumed advantages of the metaphor-based interfaces may have allowed much greater improvements in performance, with the greatest improvements occurring in the most appropriate metaphor. This trend was evident in the data, as the standard service showed no improvements in attitude, and few improvements in performance over time compared to the metaphor-based services. Overall, the office filing system service emerged as being the service that was perceived most positively, and generated the best performance levels after 3 trials.

It should be noted however that the metaphor services used were prototype designs, and were not professionally implemented services. Several suggestions for improvement emerged from post experiment interviews held with participants. Many of the criticisms were common to all services, such as: ‘the messages and prompts were too repetitive’, and ‘navigating backwards was unintuitive and difficult’. Some negative themes specific to the office filing system service did emerge: ‘it was difficult to remember colours as menu options...the assignment of colours to options seemed a bit random’; ‘you can’t anticipate options because they are different each time’; and ‘you had to listen really hard to the options when you first use it’. When participants were asked which single thing they would change about the service, the most frequently cited suggestions were: reduce dialogue length, allow menus to be skipped, and remove the different personalities (the metaphor services consisted of different role-playing characters such as the lift operator, the customer service assistant, and the store manager). The present data may not then reflect the true performance advantage of professionally developed metaphor services, and the possibility of merging some of the more positive features of each service may represent a suitable way forward for future metaphor development.

There were also important observations made by the experimenter regarding the way that different participants approached and used the services. Firstly, there were individual differences in the degree to which participants immersed themselves in the services, with some participants saying that they were merely word spotting and that

the metaphor made no difference, whilst other participants visualised the service and engaged with the different characters. Secondly, female participants tended to be more positive about the characters used in the service, saying they made it more interesting and more fun, whilst the majority of male participants either did not register them or found them unnecessary, annoying, or patronising. Finally, other participants appeared to consistently dislike particular metaphors throughout the study. Such differences in attitude and behaviour must be investigated with reference to the specific attributes of participants. This would then enable future designs to accommodate for individual differences by capitalising on users' strengths and personal preferences and characteristics.

5.6 Chapter summary

It has been demonstrated that a speech-based automated city guide service for mobile phones can be designed according to established guidelines for speech systems, can be based on an underlying metaphor, and can be usable. After a relatively low number of exposures to a metaphor-based service, user performance and attitudes were found to be equivalent to those derived from a standard menu service of equal complexity and requiring equal task involvement. The metaphor-based services also tended to be more positively evaluated at the completion of experimental testing, suggesting that such services may lead to greater user acceptance and improved performance over longer periods of use. A likely rationale for this effect is that participants may require more time to learn and accept a new metaphor-based interface, especially when the current predominant interface style is so well established.

Participants both preferred and performed better with the least complex metaphor, the office filing system, which could be both a reaction to the already high cognitive effort necessary to operate in a mobile environment, and because it is a familiar metaphor from desktop computer usage. It may be suggested that for mobile phone services, either a simpler structured metaphor is needed than more complex alternatives such as a department store (Dutton et al. 1999), or that different metaphors are needed for different service domains. Previous work, conducted in preliminary study one of this thesis has suggested that similar metaphors can be successfully applied to two different information retrieval service domains. In addition, since real world office filing systems can potentially contain information

from a diverse range of domains, then it can be hypothesised that an office filing system metaphor might be the most versatile metaphor to adapt to different mobile phone service domains.

The extent to which single metaphors might generalize to other more complex service domains cannot be concluded from this experiment. A possible restriction on the degree to which the metaphor could be applied across domains is if a service allowed the user more functionality than simply information retrieval, and a greater degree of interactivity and flexibility. In this case it may be possible that a single metaphor may not be appropriate due to the potential difficulties of scaling metaphors to cover the full system functionality of more complex systems. Under such circumstances, combined metaphors may be needed (Rumelhart and Norman, 1981).

It may be argued that the usefulness of a metaphor based upon a filing cabinet is limited, given that more and more documentation is being stored and retrieved electronically. However, it is conceivable that the components of the physical office filing system metaphor might be transformed into an electronic office filing system metaphor, although more research must be conducted to generate the correct dialogue for such a service, and to evaluate its usability compared with that of the current office filing system service.

Some participants mentioned that they liked the metaphor, but thought it would be difficult to use if they were out and about, due to background noise and visual distractions (e.g. bar, cafe, common room, bus), and also because it was more difficult to visualise the metaphor-based service structure when they could see a real world environment that was different to, and therefore clashed with what they were trying to visualise. Moreover, many participants complained of social pressures making them feel uncomfortable when saying the service menu options. In such cases, participants felt awkward or self conscious saying menu options into their telephone that would seem disjointed and nonsensical to bystanders. These could be key issues in the use of interface metaphors if they are to be used in a wide range of mobile environments. These findings lend support to the model of HCI proposed by Eason (1991), which posits 'context of use' as one of the components affecting human-computer interaction. It is for this reason that a more formal analysis of the effects of context of

use on usability was conducted in experiment two. A second component of the Eason model of HCI (1991) 'the human' was also investigated in experiment two, by measuring a range of participants' individual differences, and analysing the patterns with which they affect service usability in different contexts of use. Finally, the three days of exposure to the services during experiment one may have been enough time to measure user's learning, but not their retention about how to use the services. Experiment two extended the number of trials, and the period of time between trials, in order to measure both learning and retention.

:: CHAPTER 6

The impact of interface metaphor and context of use on the usability of a speech-based mobile city guide service

6.1 Introduction

In experiment one, a non-metaphor standard version and three metaphor-based versions (travel system, office filing system, shopping) of a speech-based mobile city guide service were implemented. After three consecutive days of usage, the office filing system metaphor service led to levels of performance and attitude that were equal to the standard service, and which tended to be better than the other metaphor-based services. It was for this reason that the office filing system metaphor was selected for further development and investigation in experiment two, which is reported in this chapter. It was hypothesised that the office filing system service was found not to be significantly better than the standard service because participants had relatively little practice with it, and that usage over a longer period of time would allow users the additional time needed to learn and retain knowledge about the service, leading to a subsequent improvement in their performance. It was for this reason that a more longitudinal experiment was designed for experiment two.

Another finding from experiment one was that the context of use in which the services were used may have been an important factor affecting the interaction. Context of use was therefore investigated as part of experiment two, specifically the physical location and the social setting, with a division into private and public locations being formulated as a means of comparing the dynamic mobile context of use with a static context of use.

Finally, the second component of the Eason (1991) model of HCI ‘the human’ was investigated. A range of participants’ individual differences were measured as part of experiment two, and their impact on interaction with the services, in both private and public contexts of use, was analysed. These individual differences, and a rationale for their inclusion, were discussed in section 2.4.1 of chapter two.

The current experiment differs from experiment one, in that, in order to examine participant’s retention of both the metaphor-based service (office filing system), and the non-metaphor standard service, the two services were evaluated at regular intervals over a longer (6-week) period, the location in which services were used was controlled, and a range of individual difference measures were included. A longitudinal approach was adopted whereby each participant’s initial level of performance could be compared with that of subsequent weeks for each service.

In relation to the non-metaphor standard service and the metaphor-based office filing system service implemented for this experiment, the primary objectives of experiment two were:

1. To compare the usability of the services over an extended period of time
2. To investigate the effect of private and public location on the usability of the services
3. To investigate the impact on usability of the following individual differences: age, gender, working memory, verbal ability, spatial ability, cognitive style, previous mobile phone experience, previous fixed line telephone experience, previous computing experience, and attitude towards mobile phone usage in public.

It was expected that the metaphor-based office filing system service would lead to improved levels of performance relative to the non-metaphor standard service, and that performance levels in public locations would be poorer than performance levels in private locations, for both services.

6.2 Prototype design and development

The size and content of the services was the same as those used for experiment one, and consisted of 5 levels of service messages and prompts, with a maximum of 3 menu options to choose from at each level. The standard service was the same as the service used in experiment one, but, on the basis of participant feedback, some changes were made to the office filing system service. Firstly, the two service characters (voices) from experiment one, the ‘office manager’, and the ‘folder information service’ were reduced to one for this experiment (office manager). Whereas in experiment one, the office manager introduced herself personally, as ‘Kate’, in this prototype she introduced herself by her job title ‘I am the office manager’. Secondly, the coloured folders at level four of the service were changed to numbered folders. Table 6.1 provides an overview of the two services, and an example of the dialogue from each level of the services

Table 6.1. Interface objects, menu options, and dialogue for the 2 services used for experiment two

Service	Level	Objects	Options	Example dialogue
Standard	1	3 options	One, two, and three	Welcome to the telephone city guide service. There are 3 options available. For Nightlife select option 1, for Eating Out select option 2, and for Arts and Entertainment select option 3. Say 1, 2, or 3.
	2	2 options	One, and two	Option 1, Nightlife. There are 2 categories of Nightlife available. For Music select option 1, and for Drinking select option 2. Say 1, or 2.
	3	3 options	One, two, and three	Option 2, Drinking. There are 3 categories of Drinking available. For Pubs select option 1, for Wine Bars select option 2, and for Pre-Club Bars select option 3. Say 1, 2, or 3.
	4	3 options	One, two, and three	Option 1, Pubs. There are 3 categories of Pubs available. For Pubs that close at 11pm select option 1, for Pubs that close at 12 midnight select option 2, and for Pubs that close at 1am select option 3. Say 1, 2, or 3.
	5	City information	No menu options	Option 2. The Pubs that close at 12 midnight are The Beach and The Quadrant.
Office filing system	1	3 filing cabinets	Left, middle, and right	Welcome to the office city guide. I am the office manager, and I will help you to find the information you want. You are at the main office. There are 3 filing cabinets available. For Nightlife information select the filing cabinet on the left, for Eating Out information select the filing cabinet in the middle, and for Arts and Entertainment information select the filing cabinet on the right. Say left, middle, or right.
	2	2 drawers	Top, and bottom	You selected the Nightlife filing cabinet. There are 2 drawers available. For Music information select the top drawer, and for Drinking information select the bottom drawer. Say top, or bottom.
	3	3 partitions	Front, middle, and rear	You selected the Drinking drawer. There are 3 partitions available. For Pub information select the front partition, for Wine Bar information select the middle partition, and for Pre-club Bar information select the rear partition. Say front, middle, or rear.
	4	3 folders	First, second, and third	You selected the Pub partition. There are 3 folders available. For Pubs that close at 11pm select the first folder, for Pubs that close at 12 midnight select the second folder, and for Pubs that close at 1am select the third folder. Say first, second, or third.
	5	City information	No menu options	You selected the second folder. The Pubs that close at 12 midnight are The Beach and The Quadrant.

6.3 Methodology

6.3.1 Design

The experimental design was a 7 x 2 x 2 mixed factorial design with 3 independent variables: 1 between-subjects factor, and 2 within-subjects factors. The between-

subjects factor was the type of service, which had 2 levels: the standard service, and the office filing system service. The first within-subjects factor was location, which consisted of 2 levels: private, and public. Participants were told that a private location would be considered as somewhere quiet, indoors, and where they were alone, for example, their bedroom, whereas a public location would be considered as somewhere outside their home and accessible to the public, for example, a café. In order to control the settings, an additional requirement was for users to be stationary and not directly engaged in other tasks. The final choice of location was determined by the participant, and so was effectively subjectively 'private' or 'public' to them. Table 6.2 shows the distribution of calls made from private and public locations with a similar range of locations being selected by participants using both services.

Table 6.2. Percentage of calls for both private and public locations for experiment two

Private location			Public location		
Location	Frequency	Percentage	Location	Frequency	Percentage
Bedroom	86	51.2	Street	67	39.9
Lounge	64	38.1	Main road	16	9.5
Classroom	7	4.2	Corridor	11	6.5
Kitchen	6	3.6	Pedestrian zone	10	6.0
Car	3	1.8	Shopping mall	10	6.0
Office	2	1.2	Shopping street	9	5.4
			Building foyer	8	4.8
			Car park	8	4.8
			Café	7	4.2
			Park	5	3.0
			Launderette	5	3.0
			Bar	4	2.4
			Supermarket	4	2.4
			Train	4	2.4

The second within-subjects factor was the trial consisting of 7 testing sessions conducted at weekly intervals. Repeated testing with the services over 7 weeks was conducted in order to establish a reliable estimate of system performance. Each participant completed a control trial using the standard service in a private location, and then 6 experimental trials using one of the 2 versions of the service. On experimental trials, participants were assigned to one of two locations (private or public), which alternated weekly. The trial location was counterbalanced across trials and between experimental groups, and although factors such as background noise could not be controlled, it may be assumed that such factors might balance out across

weeks and conditions such that their net effect would be minimised across any observed main statistical effects (Robson 2002).

6.3.2 Participants

Fifty-six people took part in the study, consisting of undergraduate students aged between 18 and 36 (mean age 21), of whom 14 were male and 42 were female. The participants were matched for age and gender, and were divided equally between the 2 experimental groups. All participants were recruited from the participant pool at the University of Portsmouth psychology department.

6.3.3 Apparatus

A WOZ methodology (Fraser and Gilbert 1991) was used for the experiment. The same experimental set up was used as in experiment one, and each of the services was designed as a single webpage, with hyperlinks linking to sound files, and the experimenter activating the sound files requested by participants. As in experiment one, participant interactions with the services were logged using TrueActive monitoring software (TrueActive Corporation, Kennewick, WA).

6.3.4 Data collection

Four types of data were collected during the study: performance measures; attitude measures; the qualitative data from the post-task telephone interviews; and finally the individual differences data. The 8 performance measures recorded were the same as those taken in experiment one. However, the majority of these measures were not analysed for the following reasons. Successful task completion was not analysed due to the high task completion rates achieved, with most participants adopting a strategy of persisting with each task until they had found the correct information. Error rates were not analysed since few errors were found across conditions. Repeat and Return rates were measured, but due to their infrequent usage, were also not analysed. The final objective measure omitted from the experiment two analysis was the 'total number of nodes used to successfully complete tasks'. It was decided to simply analyse the other node-based measure 'number of nodes used to complete tasks (as a percentage of the optimum)', as this provided information about the path through the service relative to the shortest path, and allowed tasks of different difficulty to be

compared. A total of three performance measures were therefore analysed for experiment two, and can be seen in Table 6.3 (absolute transaction times can be seen in Appendix 21). The 6 attitude measures were the same as those used for experiment one, and were collected after each trial using the same 50-item 7-point Likert scale questionnaire (see Appendix 2), with higher scores on each scale indicating a positive attitude towards the factor (see Table 6.4).

Table 6.3. Performance measures for experiment two

Measure	Description
Time	The time to complete a task was calculated as a percentage of the total length of the service dialogue (a lower score was taken to indicate a better performance)
Prompt interrupts	The number of times the dialogue was interrupted with a response was calculated as a percentage of the total number of dialogue nodes used (a higher score was taken to indicate a better performance)
Nodes	The number of dialogue nodes used to complete a task was calculated as a percentage of the optimum number – the shortest route (a lower score was taken to indicate a better performance)

Table 6.4. Attitude measures for experiment two

Measure	Description
System response accuracy	Refers to how accurately the service recognised user input, and provided a response that matched the user's expectations
Likeability	Refers to whether the user found the service useful and pleasant and whether they would choose to use it again
Cognitive demand	Refers to the user's opinion about the amount of effort necessary to use the service, and how they felt as a result of expending this effort
Annoyance	Refers to the degree to which the user found the service repetitive, boring, irritating or frustrating
Habitability	Refers to whether the user's conceptual model of the service was sufficient to inform them of what to do and what the service was doing
Speed	Refers to how quickly the user perceived the service as responding to their input, and to the overall duration of the interaction

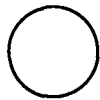
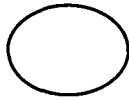

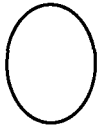
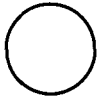
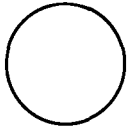
The questions used for the post-task informal telephone interviews, which were conducted after the final testing session, can be seen in Appendix 15. In terms of the individual differences data gathered, the technographic questionnaire, as used in experiment one, was used to gather data about age, gender, previous mobile phone experience, previous fixed line telephone experience, and previous computing experience (see Appendix 10).

The AH4 test was used to gather data about verbal ability and spatial ability. The AH4 Test of General Intelligence (Heim, 1970) is a two-part test that can be used to

measure both verbal and spatial ability. The rationale for the measurement of verbal and spatial ability is presented in sections 2.4.1.4 and 2.4.1.5 of chapter two. The test is designed for participants with a wide range of abilities. Each part of the test consists of 65 multiple-choice questions with 5 possible answers, and participants are allowed 10 minutes to complete each part. The questions appear in ascending order of difficulty, and participants have to complete as many questions as they can within the 10-minute time limit. Part 1 tests verbal ability, and an example of the type of question from this part is:

	1	2	3	4	5
Prevent means the same as...	avoid	cure	allow	deter	help

An example of the type of question from Part 2, which measures spatial ability, is:

	1	2	3	4	5
 is the same as...					

The computer-based Cognitive Styles Analysis (CSA) was used to gather data about cognitive style. The rationale for the measurement of cognitive style is presented in section 2.4.1.6 of chapter two. One of the most widely used instruments to measure cognitive style is Witkin's group Embedded Figures Test (GEFT) (Oltman, Raskin, Witkin, 1971). This test only measures field independence by assessing participants' ability to locate shapes embedded in more complex shapes. However, by using this approach GEFT reveals a limitation, because levels of field dependence are calculated as the result of field-independence capability. Field dependence is not, therefore, positively measured, but is inferred from poor field-independent capability. The Cognitive Styles Analysis (CSA), developed by Riding (1991), overcomes this limitation through the use of two sub-tests. The CSA measures a wholist/analytic dimension, which is equivalent to field dependence/field independence (Riding and Sadler-Smith, 1992). In the first sub-test, participants must assess the similarities of a

series of complex geometric figures, which requires the use of field-dependent capacities. In the second sub-test the participants perform a task, similar to that of the GEFT, by locating shapes embedded in more complex shapes, which requires field-independent capacities. Thus, field dependence is positively measured. The CSA was therefore judged to be a more detailed tool for the measurement of cognitive style.

The Baddeley working memory test was used to gather data about working memory. The Working Memory Span Test (Baddeley, 1986) was used to assess participant's working memory capacity, and the rationale for the measurement of working memory is presented in section 2.4.1.7 of chapter two. Participants were presented with 5 sentences, which were either semantically correct or incorrect and consisted of 5 words each, for example 'The boy brushed his teeth' and 'The train bought a newspaper'. After listening to each sentence, participants were asked to say 'yes' if the sentence made sense, and 'no' if it did not. This formed the information processing part of the test, and acted as a continuous task, which prevented rehearsal. After hearing the 5 sentences, participants were asked to recall the last word of each of the 5 sentences in the correct order. Eight blocks of sentences were used in the study, yielding a maximum test score of 40 for correctly recalled items. This part formed the memory part of the test. The test therefore required the participant to perform two tasks simultaneously: an information processing task and a memory task. The advantage of using this test is that it is conducted verbally, and places similar demands on the participant as those that are required when using a mobile phone service: the service presents lists of menu options which the user must remember; and the user must also simultaneously process information about their environment in order to successfully navigate or perform other tasks.

The mobile phone attitude questionnaire was used to gather data about participants' attitudes towards mobile phone usage in public (see Appendix 16). The mobile phone attitude questionnaire was designed to assess participants' social attitudes towards mobile phone usage in public, and a rationale for the measurement of such attitudes was presented in section 2.4.1.8 of chapter two. The questionnaire consisted of 20 statements (10 positively worded and 10 negatively worded), with a 7-point Likert response scale ranging from 'strongly disagree' to 'strongly agree'. A high score indicated a more positive attitude towards mobile phone usage in public. Salient

examples of statements from the questionnaire included: I feel awkward answering a mobile phone call when I am in a busy public place; I feel uncomfortable when people are talking on mobile phones near me in public places; I do not like the feeling of being watched when I use my mobile phone in public places. The validity of the questionnaire was established through a process of content validity, and the same procedure was followed as that described when validating the usability questionnaire.

6.3.5 Procedure

In a preliminary session, participants completed the technographic questionnaire, the AH4 test, the CSA, the working memory test, and the mobile phone attitude questionnaire. Participants were then given the experimental materials needed to complete the experiment: a participant information form outlining the experimental procedure; a response scale to refer to when verbally responding to the Likert questionnaire statements; a practice task sheet; and a task sheet containing the task to be completed during each of the 7 trials. Each task required the participants to find a specific piece of information, for example, 'Find the names of 2 pubs that close at 1am and then exit the service'. The tasks became progressively more difficult across the trials, with later tasks comprising 2 subtasks that required participants to go backwards as well as forwards through the service structure, for example, 'Find the names of 2 mid-range Italian restaurants, and then find the names of 2 pop concerts, and then exit the service'. The full task list for all trials can be seen in Appendix 17. All participants were informed that they could interrupt the service with a response at any time.

For the next part of the experiment, participants were called on their mobile phones and were required to attempt one task each week using one of the city guide services. The experimenter activated the service messages and prompts, and the participants responded verbally with their menu selections. On completion of each testing session, participants verbally responded to the Likert questionnaire statements, and were asked to identify where they were when they used the service, about any visual or auditory distractions they experienced, and whether they perceived these distractions as negatively affecting their performance. This procedure was repeated over the following 6 weeks at approximately weekly intervals. After the final testing session participants were interviewed regarding their experience of using the service.

6.4 Results

One-way ANOVA tests were used to compare differences between the 2 experimental groups for age, mobile phone experience, fixed line telephone experience, computing experience, working memory, verbal ability, spatial ability, cognitive style, and attitude towards mobile phone usage in public. No significant differences between groups were found for any of the individual difference measures. Table 6.5 show descriptive data for all individual difference measures. With respect to working memory, verbal ability, and spatial ability, a higher score indicated better performance on that test. For cognitive style, a lower score indicated a more wholist tendency, whilst a higher score indicated a more analytic tendency. For the mobile phone attitude measure, a higher score (on a scale of 1-7) indicated a more positive attitude towards mobile phone usage in public.

Table 6.5. Descriptive data for the individual difference measures for experiment two

Participant variable	Minimum	Maximum	Mean	Standard deviation
Age	18.00	36.00	20.98	4.28
Working memory	13.00	36.00	26.17	5.23
Verbal ability	19.00	57.00	40.77	8.15
Spatial ability	27.00	65.00	53.55	8.19
Cognitive style: WA ratio	0.59	2.88	1.41	0.50
Mobile phone experience	1.71	64.14	35.86	13.68
Telephone experience	0.00	70.40	31.91	17.91
Computing experience	27.25	100.00	72.67	17.01
Mobile phone attitude	2.40	6.30	4.36	0.86

All participants used the standard service for the first (control) trial. One-way ANOVA tests were used to examine whether there were any differences on the 6 attitude and 3 performance measures between experimental groups for the control trial. Only one significant difference was found for the performance measure of prompt interrupts ($F(1,54) = 5.39, p = 0.02$), with participants in the metaphor service group making more prompt interrupts than participants in the standard service group. To control for differences in users' initial level of performance or attitude towards the service, the performance and attitude scores from the control trial were used as covariates in subsequent analyses of experimental trial data. A multivariate approach was adopted as the aim of the study was to examine the impact of the metaphor-based service across several measured sources of variance (dependent variables), which taken together comprised a composite measure of usability.

Table 6.6 shows the range of possible scores for the performance and attitude measures, as well as the grand mean and standard deviation for all users and across both services.

Table 6.6. Score range and means for the performance and attitude measures for experiment two

Measure type	Measure	Range	Mean	SD
System performance measures	Time	0-100	55.80	6.99
	Prompt interrupts	0-100	97.10	3.45
	Nodes	>100	115.39	7.31
Attitude measures	System response accuracy	1-7	5.42	0.66
	Likeability	1-7	5.07	0.63
	Cognitive demand	1-7	5.43	0.68
	Annoyance	1-7	4.34	1.11
	Habitability	1-7	5.09	0.81
	Speed	1-7	4.55	1.18

6.4.1 The effect of metaphor on performance and attitude

Data from the 3 performance measures were analysed using a 6 (trial) x 2 (experimental group) MANCOVA to explore differences between the 2 experimental groups. Table 6.7 shows the means for each of the 3 performance measures across the 6 weekly trials. The data presented in Table 6.7 for each week is the data for both private and public locations. The locations were counterbalanced across weeks so that half of the participants from each experimental group used the service in private for week 1, and half of participants used the service in public for week 1. These participants then used the service in different locations for week 2, with location being alternated in subsequent weeks of testing.

Table 6.7. Descriptive statistics for performance measures across the 6 trials

Performance measure	Experimental group	Mean and SD	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
Time	Non-metaphor	Mean	65.52	59.97	56.69	58.15	59.14	56.92
		SD	9.56	9.59	10.26	6.77	9.52	9.45
	Metaphor	Mean	61.45	52.33	49.89	49.59	50.90	49.04
		SD	7.73	6.41	6.96	5.63	7.94	6.40
Prompt interrupts	Non-metaphor	Mean	89.60	95.05	96.62	97.95	97.96	97.66
		SD	10.85	8.73	6.51	4.83	4.57	5.29
	Metaphor	Mean	94.76	99.60	98.25	99.29	99.46	99.05
		SD	6.59	2.10	4.70	2.62	1.98	2.84
Nodes	Non-metaphor	Mean	113.10	106.75	110.00	108.57	133.57	129.64
		SD	16.30	8.19	9.03	10.79	18.90	28.35
	Metaphor	Mean	104.37	101.59	114.64	107.14	131.79	123.57
		SD	6.99	3.96	12.90	4.60	21.09	19.29

The multivariate effect of experimental group was significant ($F(3, 49) = 6.70, p < 0.01$, Wilks' $\lambda = 0.71$). Simple univariate tests showed that participants using the office filing system service performed tasks significantly faster ($F(1, 51) = 15.14; p < 0.01$), and interrupted prompts significantly more frequently than the standard service ($F(1, 51) = 7.15; p = 0.01$), but there was no difference in the number of nodes used ($F(1, 51) = 1.43; p = 0.24$). These differences can be seen in Figures 6.1 to 6.3. There was no multivariate effect of trial ($F(15, 37) = 0.50, p = 0.93$, Wilks' $\lambda = 0.83$), but the multivariate effect for the interaction between trial and experimental group was significant ($F(15, 37) = 1.97, p = 0.05$, Wilks' $\lambda = 0.56$). Overall, this suggests that performance with the office filing system service was better than that with the standard service, but that improvements in user performance over time did not occur at the same rate for the two service types.

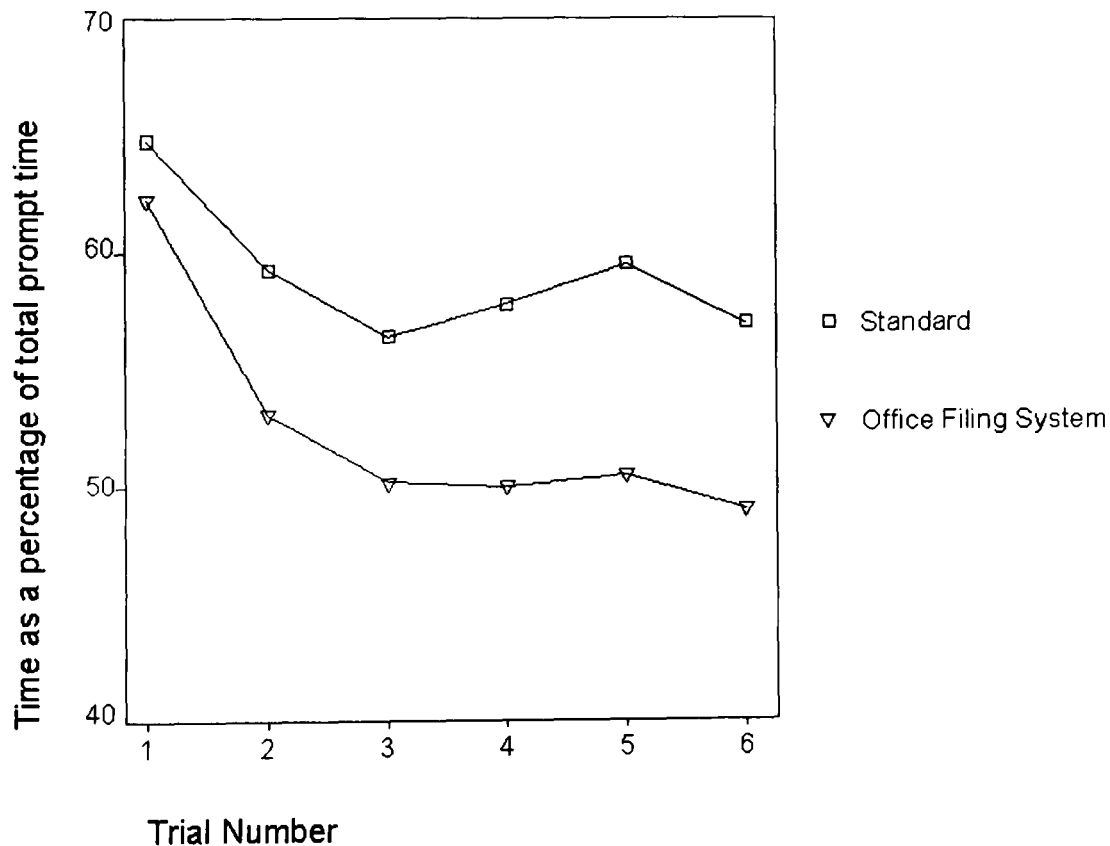


Figure 6.1. Mean scores across experimental trials for time

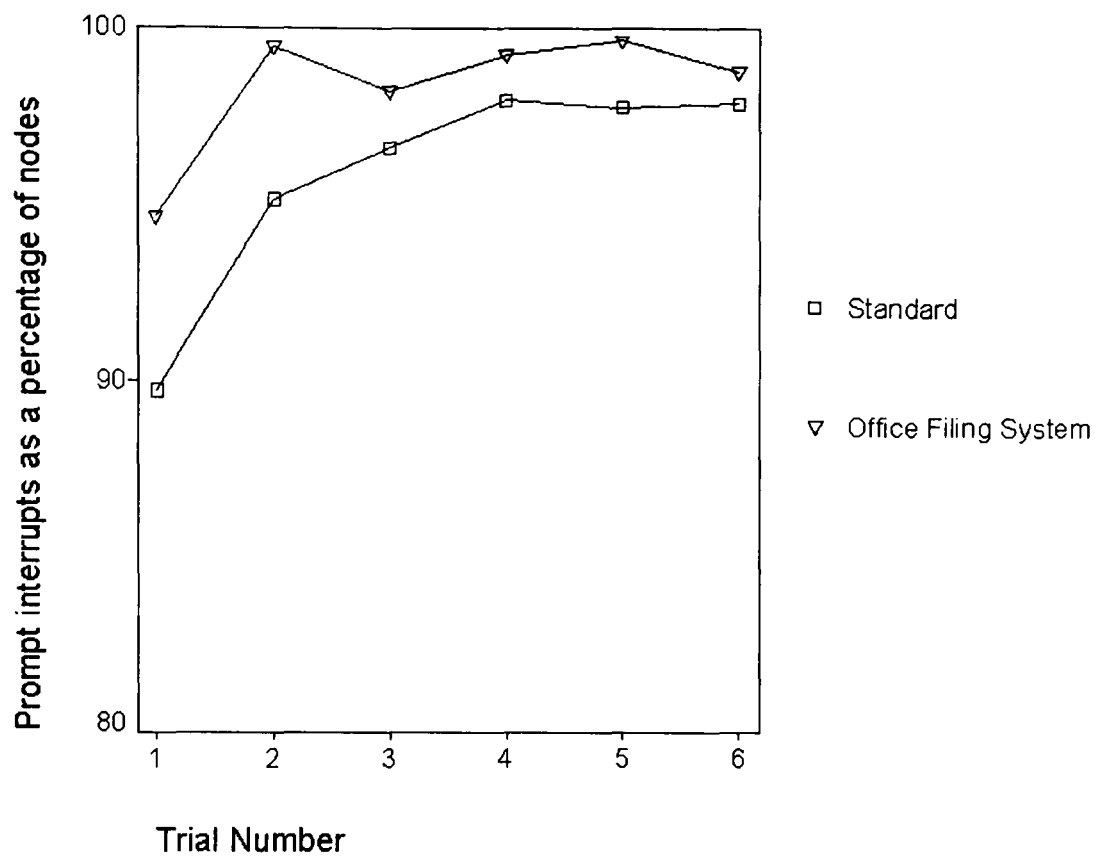


Figure 6.2. Mean scores across experimental trials for prompt interrupt

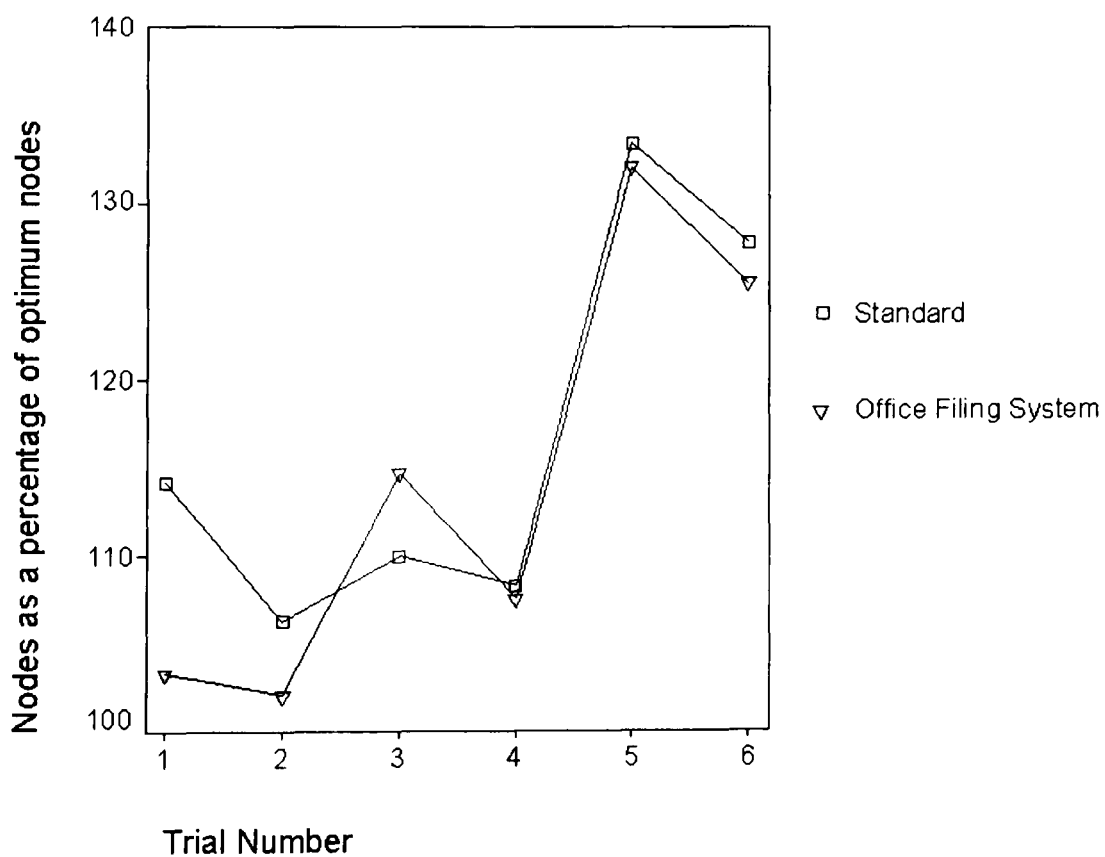


Figure 6.3. Mean scores across experimental trials for nodes

Attitude data were also analysed with a 6 (trial) x 2 (experimental group) MANCOVA to establish whether there were any differences in user preferences for the 2 services. Table 6.8 shows the means for each of the 6 attitude measures across the 6 weekly trials.

Table 6.8. Descriptive statistics for attitude measures across the 6 trials

Attitude measure	Experimental group	Mean and SD	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
System response accuracy	Non-metaphor	Mean	5.34	5.45	5.37	5.28	5.28	5.25
		SD	0.61	0.66	0.70	0.97	0.91	0.83
	Metaphor	Mean	5.42	5.67	5.41	5.78	5.27	5.50
		SD	0.82	0.70	0.82	0.69	1.13	1.14
Likeability	Non-metaphor	Mean	5.03	5.02	4.94	4.94	5.07	4.95
		SD	0.62	0.55	0.69	0.78	0.72	0.79
	Metaphor	Mean	5.12	5.15	4.96	5.30	5.09	5.25
		SD	0.69	0.64	0.75	0.60	0.80	0.76
Cognitive demand	Non-metaphor	Mean	5.46	5.48	5.46	5.41	5.38	5.25
		SD	0.62	0.71	0.68	0.98	0.74	1.00
	Metaphor	Mean	5.34	5.54	5.21	5.80	5.32	5.51
		SD	0.97	0.94	0.85	0.62	1.09	0.99
Annoyance	Non-metaphor	Mean	4.29	4.26	4.19	4.20	4.14	4.03
		SD	1.20	1.09	1.18	1.33	1.20	1.23
	Metaphor	Mean	4.20	4.61	4.30	4.69	4.51	4.67
		SD	1.24	1.17	1.29	1.13	1.25	1.23
Habitability	Non-metaphor	Mean	5.16	5.26	5.06	5.20	4.84	4.91
		SD	0.81	0.77	0.83	1.06	0.99	0.95
	Metaphor	Mean	5.12	5.36	4.80	5.61	4.70	5.02
		SD	1.17	1.03	1.03	0.80	1.30	1.34
Speed	Non-metaphor	Mean	4.31	4.57	4.38	4.13	4.29	4.09
		SD	1.13	1.12	1.23	1.28	1.40	1.32
	Metaphor	Mean	4.87	4.85	4.79	4.76	4.67	4.88
		SD	1.34	1.20	1.47	1.36	1.45	1.38

The multivariate effect of experimental group was not significant ($F(6, 43) = 0.52, p = 0.79$, Wilks' $\lambda = 0.93$). However, based on the means for all subjective variables over the 6 weeks of testing, the office filing system service was rated more positively than the standard service on 78% of the trials. There was no multivariate effect of trial ($F(30,19) = 0.64, p = 0.87$, Wilks' $\lambda = 0.50$), and no multivariate effect for the interaction between trial and experimental group ($F(30,19) = 1.06, p = 0.46$, Wilks' $\lambda = 0.38$). This suggests that overall the office filing system service was not perceived as being significantly better or worse than the standard service, and that user perceptions towards the services were consistent over time.

6.4.2 The effect of context on performance and attitude

To explore the effect of context, participant data were combined by calculating an overall mean for each of the 3 performance measures on the 3 trials conducted within each context (Note that this analysis meant combining data from consecutive weeks but that the effect of trial week was balanced across context). The data from the 3 performance measures were analysed using a 2 (context) x 2 (experimental group)

MANCOVA to explore differences (i) between the 2 experimental groups for each context (ii) between the 2 contexts within each group (see Table 6.9).

Table 6.9. Descriptive statistics for performance measures for private and public use

Performance measure	Experimental group	Private location		Public location	
		Mean	SD	Mean	SD
Time	Non-metaphor	59.83	8.19	58.97	7.08
	Metaphor	52.09	5.06	52.31	5.17
Prompts interrupts	Non-metaphor	94.90	6.13	96.72	4.18
	Metaphor	98.22	2.85	98.58	2.30
Nodes	Non-metaphor	119.14	11.44	114.74	10.03
	Metaphor	115.94	9.32	111.76	7.59

The multivariate interaction between context and experimental code was not significant ($F(3, 49) = 0.51, p = 0.68, \text{Wilks}' \lambda = 0.97$). Multivariate main effects tests for each of the 2 contexts of use showed a significant effect for both private location ($F(3, 49) = 4.74, p = 0.01, \text{Wilks}' \lambda = 0.78$) and public location ($F(3, 49) = 4.78, p = 0.01, \text{Wilks}' \lambda = 0.77$), which arose since participants using the office filing system service performed faster than the standard service in both private ($F(1, 51) = 14.19, p < 0.01$) and public ($F(1, 51) = 11.50, p < 0.01$) locations, and used significantly more prompt interrupts when in a private location ($F(1, 51) = 6.17, p = 0.02$) than did the participants using the standard service. Performance on all 3 measures in both private and public were better with the office filing system service than with the standard service. Multivariate main effects tests showed no significant differences in performance between private and public usage for the standard service ($F(3, 49) = 1.99, p = 0.13, \text{Wilks}' \lambda = 0.89$) or the office filing system ($F(3, 49) = 1.08, p = 0.37, \text{Wilks}' \lambda = 0.94$).

A 2x2 MANCOVA was also used to explore the interactions between experimental group and context for user attitude measures (Table 6.10).

Table 6.10. Descriptive statistics for attitude measures for private and public use

Attitude measure	Experimental group	Private location		Public location	
		Mean	SD	Mean	SD
System response accuracy	Non-metaphor	5.34	0.66	5.31	0.70
	Metaphor	5.41	0.85	5.60	0.56
Likeability	Non-metaphor	4.98	0.67	5.00	0.63
	Metaphor	5.09	0.73	5.20	0.53
Cognitive demand	Non-metaphor	5.45	0.70	5.37	0.66
	Metaphor	5.40	0.90	5.51	0.65
Annoyance	Non-metaphor	4.18	1.15	4.19	1.16
	Metaphor	4.44	1.19	4.56	1.06
Habitability	Non-metaphor	5.06	0.77	5.08	0.74
	Metaphor	5.02	1.10	5.17	0.81
Speed	Non-metaphor	4.26	1.13	4.33	1.12
	Metaphor	4.76	1.29	4.84	1.19

The multivariate interaction between context and experimental group was not significant ($F(6, 43) = 0.57, p = 0.75, \text{Wilks' } \lambda = 0.93$). Multivariate simple main effects tests for each of the 2 contexts of use revealed no significant differences in attitude between services for private usage ($F(6, 43) = 0.53, p = 0.78, \text{Wilks' } \lambda = 0.93$) or public usage ($F(6, 43) = 0.68, p = 0.67, \text{Wilks' } \lambda = 0.91$). However, the office filing system service was, on average, rated more positively than the standard service when used in private for 4 of the 6 measures, and when used in public for all 6 attitude measures. Multivariate main effects tests showed no significant differences in attitude between private and public usage for the standard service ($F(6, 43) = 0.73, p = 0.63, \text{Wilks' } \lambda = 0.91$) or the office filing system service ($F(6, 43) = 1.24, p = 0.31, \text{Wilks' } \lambda = 0.85$).

6.4.3 Post task interview

As part of the post task interview, participants were asked whether they had tried to visualise the service structure whilst attempting the tasks. If participants claimed to have visualised the service, their descriptions of the structure and features of the visualisation were analysed to ensure that it met the following criteria for visualisation, which were derived from the definition of visualisation ability proposed by Ekstrom et al. (1976: p. 173):

- Manipulation or transformation of a mental image of spatial patterns into other arrangements

- Mentally restructuring a mental image into components for manipulation, rather than manipulating the whole figure

Of the participants who used the office filing system service, 66% reported that they visualised the service structure. In contrast, none of the participants who used the standard structure reported any kind of visualisation beyond actually visualising the numbers that corresponded to the menu options. When participants were asked to compare the office filing system service with the standard service, the most frequently given reason for preference of the office filing system service was that they could visualise the service, which made it easier to use.

With respect to the office filing system service, there were a number of other themes that emerged from the interviews. Participants liked the clear, coherent structure of the service, which helped them to know where they were within the service, and to navigate, especially backwards through the service. Although participants were able to remember the overall service structure from week to week, the specific menu options were not very well remembered. A possible reason for this, suggested by participants, may have been the fact that the menu options were different at each level, which meant that they could not predict them as easily as they could with the numbered menu options of the standard service. Despite this, participants generally found the different menu options at different levels of the service to be helpful navigation cues, informing them that they had changed levels. Finally, when using the service in public, participants often felt uncomfortable or self-conscious saying the menu options that were required to progress through the service.

6.4.4 Multiple regression models

6.4.4.1 Performance measures

In order to investigate the association between performance, and the measured individual differences, multiple regression analyses were performed. The performance measures of interest were performance in private locations, and performance in public locations for both the standard service, and the office filing system service. The 3 performance measures analysed for this experiment (time, prompts, nodes) were recorded using different scales with different ranges. In order to reduce these 3

measures into one performance mean score for private, and one for public, it was necessary to transform the measures onto a common scale in order to make them comparable. Individual participant raw scores on each measure were therefore converted into z-scores and an overall performance mean for each participant was calculated by taking a mean of the 3 z-scores for each context of use.

The overall mean score for performance in private locations for the standard service will be referred to as standard private performance, the overall mean score for performance in public locations for the standard service will be referred to as standard public performance, the overall mean score for performance in private for the office filing system service will be referred to as office private performance, and the overall mean score for performance in public for the office filing system service will be referred to as office public performance. In order to investigate the association between the measured individual differences, and performance in private, and public locations for each service, four multiple regression analyses were performed. The criterion variables used in successive analyses were (1) the standard private performance score (2) the standard public performance score (3) the office private performance score (4) and the office public performance score. The predictor variables used for each multiple regression were age, gender, previous mobile phone experience, previous fixed line telephone experience, previous computing experience, working memory, verbal ability, spatial ability, cognitive style, and attitude towards mobile phone usage in public.

For standard private performance, the overall model was not significant ($F(10,17)=1.57$; $p=0.20$), and the adjusted R^2 value of 0.18 showed that the predictor variables accounted for 18% of the variation in the standard private performance score. None of the individual predictor variables were significantly related to the performance score. Table 6.11 shows the significance levels for all of the predictor variables.

Table 6.11. Predictors of private performance for the standard service

Predictor variables	Standardised Beta Coefficient	t	Sig.
Age	-0.39	-1.83	0.09
Gender	0.002	0.01	0.99
Working memory	-0.19	-0.83	0.42
Verbal ability	-0.07	-0.25	0.81
Spatial ability	0.47	1.45	0.17
Cognitive style	0.29	1.09	0.29
Mobile experience	-0.15	-0.70	0.50
Telephone experience	0.02	0.10	0.93
Computing experience	-0.09	-0.41	0.69
Mean mobile attitude	-0.01	-0.07	0.95

For standard public performance, the overall model was not significant ($F(10,17)=2.39$; $p=0.06$), and the adjusted R^2 value of 0.34 showed that the predictor variables accounted for 34% of the variation in the standard public performance score. Gender was found to be a significant predictor of standard public performance ($p=0.03$), but the remaining individual predictor variables did not account for a significant variance in the public performance score. Table 6.12 shows the significance levels for all of the predictor variables.

Table 6.12. Predictors of public performance for the standard service

Predictor variables	Standardised Beta Coefficient	t	Sig.
Age	0.11	0.60	0.56
Gender	-0.45	-2.31	0.03
Working memory	-0.20	-0.98	0.34
Verbal ability	-0.12	-0.47	0.64
Spatial ability	0.60	2.08	0.05
Cognitive style	-0.09	-0.39	0.70
Mobile experience	0.07	0.38	0.71
Telephone experience	0.29	1.44	0.17
Computing experience	-0.08	-0.39	0.70
Mean mobile attitude	0.25	1.36	0.19

For office private performance, the overall model was not significant ($F(10,14)=1.38$; $p=0.28$), and the adjusted R^2 value of 0.13 showed that the predictor variables accounted for 13% of the variation in the performance score. None of the individual predictor variables were significantly related to the office private performance score. Table 6.13 shows the significance levels for all of the predictor variables.

Table 6.13. Predictors of private performance for the office filing system service

Predictor variables	Standardised Beta Coefficient	t	Sig.
Age	0.44	1.82	0.09
Gender	0.10	0.37	0.72
Working memory	0.24	0.90	0.39
Verbal ability	-0.09	-0.29	0.78
Spatial ability	0.51	1.73	0.11
Cognitive style	-0.03	-0.11	0.92
Mobile experience	-0.26	-1.00	0.33
Telephone experience	-0.36	-1.46	0.17
Computing experience	0.08	0.24	0.82
Mean mobile attitude	-0.03	-0.16	0.88

For office public performance the overall model was not significant ($F(10,14)=1.07$; $p=0.44$), and the adjusted R^2 value of 0.03 showed that the predictor variables accounted for 3% of the variation in the performance score. None of the individual predictor variables were significantly related to the office public performance score. Table 6.14 shows the significance levels for all of the predictor variables.

Table 6.14. Predictors of public performance for the office filing system service

Predictor variables	Standardised Beta Coefficient	t	Sig.
Age	0.10	0.37	0.72
Gender	-0.41	-1.49	0.16
Working memory	0.33	1.16	0.26
Verbal ability	0.34	1.04	0.32
Spatial ability	-0.33	-1.05	0.31
Cognitive style	-0.37	-1.36	0.20
Mobile experience	0.46	1.64	0.12
Telephone experience	0.29	1.13	0.28
Computing experience	-0.16	-0.46	0.65
Mean mobile attitude	0.27	1.17	0.26

6.4.4.2 Attitude measures

In order to investigate the association between the attitude measures and the measured individual differences, a further set of multiple regression analyses were performed. The attitude measures of interest were attitude in private locations, and attitude in public locations for both the standard service, and the office filing system service. These scores were calculated by taking a mean of the 6 attitude measures for each participant, resulting in a score of between 1 and 7. The overall mean score for attitude towards the standard service in private locations will be referred to as standard private attitude, the overall mean score for attitude towards the standard service in public locations will be referred to as standard public attitude, the overall mean score towards the office filing system service in private locations will be referred to as office private attitude, and the overall mean score for attitude towards the office filing system service in public locations will be referred to as office public attitude.

Four multiple regression analyses were performed. The criterion variables used in successive analyses were (1) the standard private attitude score (2) the standard public attitude score (3) the office private attitude score (4) the office public attitude score. The predictor variables used for each multiple regression were the same as those used for the multiple regressions performed for the performance measures.

For standard private attitude, the overall model was significant ($F(10,17)=3.07$; $p=0.02$), and the adjusted R^2 value of 0.43 showed that the predictor variables accounted for 43% of the variation in the attitude score. Gender ($p=0.01$), working memory ($p=0.04$), and previous telephone experience ($p=0.03$) were found to be significant predictors of attitude, but the remaining predictor variables did not account for a significant variance in the standard private attitude score. Table 6.15 shows the significance levels for all of the predictor variables.

Table 6.15. Predictors of private attitude for the standard service

Predictor variables	Standardised Beta Coefficient	t	Sig.
Age	0.37	2.07	0.05
Gender	0.51	2.86	0.01
Working memory	-0.43	-2.24	0.04
Verbal ability	0.13	0.58	0.57
Spatial ability	0.40	1.51	0.15
Cognitive style	-0.30	-1.35	0.19
Mobile experience	-0.02	-0.13	0.90
Telephone experience	-0.43	-2.32	0.03
Computing experience	-0.05	-0.29	0.78
Mean mobile attitude	0.02	0.09	0.93

For standard public attitude, the overall model was significant ($F(10,17)=5.35$; $p<0.01$), and the adjusted R^2 value of 0.62 showed that the predictor variables accounted for 62% of the variation in the attitude score. Age ($p=0.03$), gender ($p<0.01$), working memory ($p=0.01$), and previous telephone experience ($p=0.01$) were found to be significant predictors of attitude, but the remaining predictor variables did not account for a significant variance in the standard public attitude score. Table 6.16 shows the significance levels for all of the predictor variables.

Table 6.16. Predictors of public attitude for the standard service

Predictor variables	Standardised Beta Coefficient	t	Sig.
Age	0.35	2.42	0.03
Gender	0.52	3.55	0.002
Working memory	-0.48	-3.07	0.01
Verbal ability	0.17	0.91	0.38
Spatial ability	0.26	1.17	0.26
Cognitive style	-0.17	-0.92	0.37
Mobile experience	0.06	0.41	0.69
Telephone experience	-0.44	-2.92	0.01
Computing experience	-0.19	-1.25	0.23
Mean mobile attitude	0.01	0.05	0.96

For office private attitude the overall model was not significant ($F(10,14)=1.04$; $p=0.46$), and the adjusted R^2 value of 0.02 showed that the predictor variables accounted for 2% of the variation in the private attitude score. None of the individual predictor variables were significantly related to the office private attitude score. Table 6.17 shows the significance levels for all of the predictor variables.

Table 6.17. Predictors of private attitude for the office filing system service

Predictor variables	Standardised Beta Coefficient	t	Sig.
Age	-0.07	-0.29	0.78
Gender	0.32	1.17	0.26
Working memory	-0.31	-1.08	0.30
Verbal ability	0.20	0.62	0.55
Spatial ability	-0.09	-0.27	0.79
Cognitive style	0.01	0.04	0.97
Mobile experience	-0.47	-1.66	0.12
Telephone experience	-0.14	-0.55	0.59
Computing experience	-0.07	-0.20	0.84
Mean mobile attitude	0.28	1.24	0.24

For office public attitude, the overall model was not significant ($F(10,14)=1.16$; $p=0.39$), and the adjusted R^2 value of 0.06 showed that the predictor variables accounted for 6% of the variation in the attitude score. None of the individual predictor variables were significantly related to the office public attitude score. Table 6.18 shows the significance levels for all of the predictor variables.

Table 6.18. Predictors of public attitude for the office filing system service

Predictor variables	Standardised Beta Coefficient	t	Sig.
Age	-0.13	-0.50	0.62
Gender	0.13	0.48	0.64
Working memory	-0.22	-0.80	0.44
Verbal ability	0.51	1.61	0.13
Spatial ability	-0.43	-1.38	0.19
Cognitive style	0.05	0.17	0.87
Mobile experience	-0.37	-1.33	0.21
Telephone experience	0.06	0.23	0.83
Computing experience	-0.31	-0.91	0.38
Mean mobile attitude	0.23	1.03	0.32

6.4.4.3 Multiple regression summary

Table 6.19 provides a summary of the individual differences that were found to be significant predictors of performance with, and attitude towards the standard service and the office filing system service, in private locations, and in public locations.

Table 6.19. Summary of significant predictors of performance and attitude

Service type	Location	Attitude	Performance
Standard service	Private	Gender, Working memory, Telephone experience	<i>No predictors</i>
	Public	Age, Gender, Working memory, Telephone experience	Gender
Office filing system service	Private	<i>No predictors</i>	<i>No predictors</i>
	Public	<i>No predictors</i>	<i>No predictors</i>

The individual difference measures recorded were not predictors of performance or attitude with the office filing system service. Gender, working memory, and telephone experience were predictors of attitude with the standard service. Age emerged as a predictor of improved perceptions towards the standard service in public locations, but was not a predictor of attitude in private locations. However, due to the narrow spread of ages across the participant sample, the effect of age will not be discussed as an important predictor. Finally, gender was a predictor of performance with the standard service when used in public, but not when used in private.

6.5 Discussion

The use of an office filing system metaphor significantly improved overall performance with the mobile city guide service in both private and public compared to a non-metaphor standard version of the service, but there were no differences in user preferences towards the two services. Tasks were performed faster with the office filing system service as a result of users interrupting the service dialogue with their menu selections more often, and more quickly. There were no differences between services in the number of information nodes visited when performing tasks. This finding suggests that when using the standard service participants were unsure about which option to choose and tended to listen more carefully to full service messages without interrupting, rather than choose an incorrect option which would have increased their node count. No significant differences in participants' initial response

to each service were evident in the current experiment, although during the first week of testing the office filing system service demonstrated improved levels of performance relative to the standard service on all measures, and more positive attitudes for half of the attitude measures. User attitude is an important factor in the acceptance of a novel service such as the office filing system service, since users who perceive the service as being hard to use at first contact may be less likely to want to use it again (Davis, Bagozzi, and Warshaw, 1989).

Users of the office filing system service interrupted the service dialogue more often than those using the standard service in private, but not in public, which may have occurred due to social inhibition. When asked at the post-task interview 'How did you feel when using the service in a public place?' participants who used the office filing system service often reported feeling slightly embarrassed or self-conscious when saying some of the menu options that were required to progress through the service, such as 'left cabinet' or 'top drawer'. Fewer participants using the standard service reported feeling inhibited, possibly since the use of numbered menu options is widespread and would not have been perceived as unusual to bystanders. When using a service in public the presence of other people may then influence how the user feels saying abstract phrases in a context where they could be overheard. Similar findings have been reported by Love and Perry (2004). Despite this, the context in which the services were used did not appear to influence performance with or preference for either of the services. This finding was corroborated by subjective reports (post-task interview). When asked about the presence of auditory and visual distractions participants usually reported at least one source of distraction, most commonly traffic, and in some cases more than one source (e.g. people talking). However, when asked whether they considered that their performance had been adversely affected they replied that it had not on at least 95% of trials. This result may be related to the fact that the majority of phone calls made by mobile phone users take place in distracting environments, such as on streets, on public transport, and in shops and restaurants (Wei and Leung 1999). Therefore, participants may have developed strategies to deal with distractions in these mobile public settings, suggesting that metaphor-based services may be used in a diverse range of settings without a significant reduction in usability.

Another interesting finding was that the majority of participants using the office filing system service reported that they visualised the service structure, but their responses suggested that this was done in different ways and with different levels of detail. When asked which service they preferred, the most commonly given reason by participants for preferring the office filing system service was that they could visualise it, and that this made it easier to use rather than just remembering numbers (which was the main strategy used for the standard service). The participants who did not visualise the service reported using either 'word spotting' or remembering sequences of menu options. It is likely that the spatial metaphor used in this study provided users with a mental model of the system in the form of a visualisation (Card et al. 1999), something that is difficult to extract from the standard service. This visualisation consisted of spatial or way-finding cues, which may have prevented users from becoming lost or confused compared to systems relying solely on sequential verbal cues. Consequently system metaphors which invoke visual images may offer a viable method of improving user interactions and may address the problems of navigation difficulties and poor usability which have been commonly associated with speech systems (Rosson 1985). The navigation of services with menu hierarchies deeper than the recommended maximum of three levels (Bond and Camack 1999) may therefore be supported, although the selection of unambiguous titles for the additional levels of the service hierarchy must remain a priority to avoid incorrect menu selections and subsequent user disorientation.

The extent to which participants' individual differences could be used to predict both attitudes towards, and performance with the services, was affected by the service type, and the location in which the services were used. In terms of service type, there were no significant predictors of attitude or performance for the office filing system service, whereas, for the standard service there were a range of predictor variables. It may have been expected that spatial ability would emerge as a predictor of either attitude, or performance, for the office filing system service. The office filing system service encouraged participants to form a spatial mental model of the service structure, and two thirds of participants reported visualising such a mental model, but despite this, spatial ability was not an important factor in how this model was utilised. It may therefore be the case that one of the sub-components of spatial ability, visualisation ability, is more salient to the mental models constructed.

Gender, working memory, and previous telephone experience were predictors of attitude towards the standard service in both private and public locations. Female participants perceived the services more positively, as did participants with low working memory, and participants who were infrequent users of fixed line telephones. These results suggest a persistent trend for male participants to dislike the standard service, regardless of location. The association between working memory and attitude was unexpected, as it may have been predicted that participants with good working memory would have perceived the services more positively due to their enhanced ability to remember the service messages and prompts. A possible explanation may be that participants with high working memory found the service easy to use, and found the menu options easy to remember, but as a result of this focussed on other aspects of the service, such as the actual user experience. They may not have liked the service, or found it tedious to use, and therefore, despite having no difficulties completing the tasks with the service, perceived it negatively. The results for previous telephone experience may relate to participant's experience with telephone-based interaction, and with automated phone services. Participants who use telephones infrequently may lack confidence in using the telephone, and prefer a rigid structured style of interaction that is task-oriented, which matches the type of limited interaction offered by the standard service.

There were no significant predictors of performance with the standard service in private locations, but gender was a predictor of performance in public locations, with males performing better than females. This suggests that context of use interacts with participant gender to affect performance with the standard service. When the standard service was used in private, there was no gender related association between the attitude and performance measures, and the poorer perceptions towards the standard service by males were not associated with any subsequent performance detriment. However, in public, there appears to have been an association between the effects of gender, and attitude and performance. When used in public, males perceived the standard service more negatively than females, but their performance was better than female performance. That this association existed in public, but not in private, may be the result of the additional physical distractions and social pressures when interacting in public, which male participants may have been better able to effectively deal with, leading to better levels of performance.

The multiple regression analyses suggest that, as none of the individual differences measured could be used to predict attitude or performance for the office filing system, the design of metaphor-based phone services may help to accommodate individual differences, making them easier to use for a wider range of people.

The current experiment extends the findings of experiment one in three main areas. Firstly, the performance benefits of a metaphor-based service become more evident over longer periods of use, but start on first exposure. Secondly, the performance benefits of a metaphor-based service are not impacted by private or public locations of use, suggesting that the visualisation strategy adopted by the majority of participants may still be viable in conjunction with everyday physical navigation in the real world. However, due to the range of possible locations within both private and public contexts, it is possible that participants may be more likely or better able to visualise the service in certain locations. Furthermore, social inhibitions were apparent in the use of the metaphor-based service dialogues as a result of the novel syntax of the menu options, but this effect may be ameliorated if service use became more widespread or through further investigations of context specific use in order to determine alternative, more acceptable phrases. Thirdly, gender, working memory, and previous fixed line telephone experience, are all related to either attitude towards or performance with the standard service, but there were no individual difference predictors for the office filing system service. Moreover, the location from which the standard service was accessed, affected the pattern with which gender could be used as a predictor of performance.

6.6 Chapter summary

User performance with a speech-based mobile city guide service improved when the service was based on an underlying metaphor. This performance advantage is consistent when the service is used in both private and public locations. Although participants did not show any significant preferences for the metaphor-based service, it was generally perceived more positively when used in public, whilst performance for first time usage also tended to be better. Social context appeared to have an effect on participants' feelings whilst using the metaphor-based service in public, but this did not adversely affect their performance when compared to private use or use of the

standard service. No conclusions can be drawn as to whether a spatial metaphor is better than other types of metaphor for mobile phone services, only that the spatial metaphor used improved performance with a mobile city guide service relative to a non-metaphor service.

The main rationale for utilising a spatial metaphor for a speech-based service was to provide a mental model of the service which could be visualised, and would provide structural cues to make the service easier to remember and easier to navigate. This appears to have been successful, with two-thirds of participants visualising the service structure, and it was this feature that was most often cited as a reason for user preference of the metaphor-based service over the standard service. In addition, a visualisation strategy did not disrupt performance with the metaphor-based service when used in public locations, which might be considered to be more visually distracting and cognitively demanding. Neither attitude nor performance could be predicted by any of the individual differences measured, highlighting the potential for metaphor-based systems to reduce the negative effects on interaction that may be caused by some individual differences. However, due to the majority of participants reporting that they visualised the office filing system service, an investigation of visualisation was conducted in experiment three.

:: CHAPTER 7

The effect of visualisation on the usability of voice-operated metaphor-based mobile phone services

7.1 Introduction

Experiments one and two have indicated that the use of interface metaphors may provide a suitable strategy for the design of hierarchically structured speech-activated mobile phone services. In experiment two, an office filing system metaphor-based version of the city guide service resulted in significantly improved user performance and was preferred to the standard number-based menu service. This performance advantage was found to be consistent over an extended 6-week usage period, and was independent of whether the context of phone use was a public or private location. This suggests that participants using the metaphor-based service were better able to retain structural information and to develop an appropriate mental model of the service than those who used the standard service.

User evaluations from experiment two suggested that the success of the metaphor may be dependent on a participant's ability to visualise its intended referent, with visualisation of the office filing system service being the most frequently cited reason for its preference over the standard service. It was for these reasons that experiment

three, which is reported in this chapter, sought to gather qualitative data to develop a deeper level of understanding about the strategies used by participants when interacting with the services, specifically the visualisation strategy.

Experiment 3 investigated visualisation of a speech-based automated mobile city guide service. The non-metaphor standard service, and the metaphor-based office filing system service from experiment two were evaluated. In addition, a spatial metaphor from a computing domain, the Microsoft Windows desktop metaphor, was used to implement a third version of the service. It was decided to implement and evaluate a service based on a computer desktop metaphor for three main reasons. Firstly, to test the effectiveness of a GUI metaphor for a speech-based system. Secondly, to compare the usability of a service based on a metaphor that was not derived from a HCD process, with the usability of the office filing system service, which was derived from a HCD process. Thirdly, to test a finding by Yankelovich et al. (1995), who discovered that the vocabulary used to describe a GUI does not transfer well to a speech-based interface because users are not in the habit of using such terminology in everyday conversation.

In experiment two, the public contexts in which the services were used varied, and therefore could not be entirely known. In this experiment context of use was controlled by selecting a context of use that was public, but similar for all participants. In experiment two, there was no significant change in performance over the 6 weekly trials, and therefore, as the focus of experiment three was visualisation, only one trial was conducted to determine participants' performance with the services on their first time use. This may also more accurately reflect real world usage patterns for such services, which increases the ecological validity of the results obtained.

The three main objectives of experiment three were to:

1. Understand the user process of visualisation, and the ways in which it may affect interaction with the services.
2. Analyse the effects of visualisation on attitudes towards, performance with, and recall of the phone services

3. Evaluate the relevance of a GUI metaphor (computer desktop metaphor) to a speech-based automated mobile city guide service.

7.2 Prototype design and development

The standard service, and the office filing system service were the same as those used in experiment two, whilst the computer desktop service was designed specifically for the purposes of this experiment. To translate the Windows desktop GUI into an auditory version of the interface, it was necessary to deal with a number of design issues. Firstly, it was necessary to decide at which level of abstraction to decompose the GUI (Mynatt, 1993; Mynatt, 1995): lexical, syntactic, or semantic. The lexical level of a GUI is the level of lines, dots, and text that make up the application interface. The syntactic level consists of lexical constructs combined into objects that convey meaning, for example buttons and menus (Edwards and Stockton, 1995). The semantic level refers to the operations that the syntactic objects allow the user to perform; for example, buttons allow the user to execute a command, whilst menus provide a list of possible commands. Translating the interface at the lexical level would not make sense to users, as individual lexical constructs do not have affordances, or 'meaning'. Translating the interface at the semantic level was not applicable to the current experiment, as the only action that may be performed upon the syntactic objects at each level of the service is to select them. It was therefore decided that the appropriate level of translation was the syntactic level, which is the level of interface components.

The next step was to identify the interface objects that directly contribute to a user's mental model of a GUI, rather than objects that are simply artefacts of visual presentation. In this way, the underlying conceptual model of a GUI could be transferred to the auditory interface. An informal evaluation was performed to investigate which interface objects were salient features of users' mental models. Five postgraduate Computer Science students were asked to verbally explain how to navigate from the desktop to a file in their personal directory. The explanations were recorded, and analysed to extract terminology concerning interface objects. Having isolated a series of objects associated with different levels of the desktop hierarchy, the objects were mapped onto the 5-level hierarchy of the phone service.

At level 1 of the service, the user was told that that they were at the computer desktop and had to choose from 2 windows. In order to navigate to the venue information contained in one of the 54 files in level 5, the user had to navigate through levels 2, 3, and 4, which involved selecting icons, folders, and files, for example *'You selected the Nightlife window. There are 2 icons available. For information about Music, select the first icon; and for information about Drinking, select the second icon. Say first or second.'*

Due to the spatial nature of the desktop interface, the role of space as a means of conveying meaning had to be considered. For the office filing system service the position of the interface objects in space were defined. However, for the computer desktop service the interface objects were not assigned to a fixed location, but it was expected that users would visualise the locations of the interface objects. This gave participants the opportunity to mentally customise the service, based on the way in which they might normally organise the various interface objects on their own computer desktops, so that the resulting system image was personal and meaningful to them.

7.3 Methodology

7.3.1 Design

The experiment used a 3 (service) x 3 (task) mixed factorial design. The between-subjects factor was the version of the service used, with participants assigned to either the standard, the office filing system, or the computer desktop service. The within-subjects factor was the task, with each participant attempting 3 consecutive tasks with their allocated service.

7.3.2 Participants

Forty-two people took part in the study, consisting of both undergraduate and postgraduate students recruited from within a university Computer Science department. Participants were aged between 18 and 30, (mean age = 22), of whom 33 were male and 9 were female. Participants were allocated equally by gender between experimental groups, such that 11 males and 3 females were in each group. All participants were highly computer-literate.

7.3.3 Apparatus

A WOZ methodology (Fraser and Gilbert, 1991) was used in this experiment. The same experimental set up was used, as in experiment two, with the experimenter selecting hyperlinks, which played the requested sound files via a speakerphone to the participant at the other end of the phone line. As in experiment two, participant interactions with the services were logged using TrueActive monitoring software (TrueActive Corporation, Kennewick, WA).

7.3.4 Data collection

Four types of data were collected during the study: performance measures; attitude measures; the qualitative data from the post-task interviews, and the retrospective analysis; and the service recall data. A total of 8 measures of performance were recorded, and were the same as those taken in experiment two. However, as in experiment two, some of these measures were not analysed for the following reasons. Return rates were not analysed due to their infrequent usage, and error rates were not analysed, since very few errors were found across conditions. In addition, as in experiment two, only one of the node-based measures was analysed, for the reasons outlined in the previous chapter. A total of five performance measures were therefore analysed for experiment three, and can be seen in Table 7.1 (absolute transaction times can be seen in Appendix 21). The 6 attitude measures were the same as those used for experiment two (see Table 7.2), and were collected using a usability questionnaire (Appendix 2).

Table 7.1. Performance measures for experiment three

Measure	Description
Time	The time to complete a task was calculated as a percentage of the total length of the service dialogue (a lower score was taken to indicate a better performance)
Task completion	The number of times the tasks were successfully completed as a percentage of the total tasks (a higher score was taken to indicate a better performance)
Prompt interrupts	The number of times the dialogue was interrupted with a response was calculated as a percentage of the total number of dialogue nodes used (a higher score was taken to indicate a better performance)
Nodes	The number of dialogue nodes used to complete a task was calculated as a percentage of the optimum number – the shortest route (a lower score was taken to indicate a better performance)
Repeats	The number of times a request to repeat a dialogue prompt was made (a lower score was taken to indicate a better performance)

Table 7.2. Attitude measures for experiment three

Measure	Description
System response accuracy	Refers to how accurately the service recognised user input, and provided a response that matched the users expectations
Likeability	Refers to whether the user found the service useful and pleasant and whether they would choose to use it again
Cognitive demand	Refers to the user's opinion about the amount of effort necessary to use the service, and how they felt as a result of expending this effort
Annoyance	Refers to the degree to which the user found the service repetitive, boring, irritating or frustrating
Habitability	Refers to whether the user's conceptual model of the service was sufficient to inform them of what to do and what the service was doing
Speed	Refers to how quickly the user perceived the service as responding to their input, and to the overall duration of the interaction

Qualitative data was gathered from the post task interviews, and the full list of questions that were used for the interviews can be seen in Appendix 18. Finally, in order to gather data about participants memory of the overall service structure, and the menu options at each level, a multiple-choice questionnaire was designed, and which can be seen in Appendix 19.

7.3.5 Procedure

At the beginning of the testing session, participants were required to complete a Technographic questionnaire (Appendix 10). Prior to performing the phone tasks, participants were given a brief explanation of the phone service they would be using, and were informed that all responses had to be spoken rather than keypad activated, and that service messages could be interrupted with a response at any time. They were then given a task sheet containing a practice task and 3 experimental tasks to be performed (Appendix 20). Each task required participants to find the names of specific venues within a fictitious city. The third task was more difficult than tasks 1 and 2, and comprised 2 subtasks that required participants to go backwards as well as forwards through the service structure.

Once the nature of the tasks had been explained, participants who consented to take part were taken to the foyer area of a University building, where they were to use the services. This area was a public setting with an always-present receptionist, and a regular flow of people entering and leaving the building. The participants were called on their mobile phones by the experimenter, and were required to attempt the practice task, followed by the 3 experimental tasks. The experimenter activated the service messages and prompts, and the participants responded verbally with their menu

selections. All participant interactions with the services were audio recorded. After completing the 3 tasks participants returned to the original location to rejoin the experimenter and complete the usability questionnaire and the multiple choice memory questionnaire requiring them to recall both the overall service structure and the menu options at each level.

At the end of the testing session, participants were asked a series of open-ended questions about their experiences of using the services. These questions could be asked in any order, and included: Were you able to visualise the structure of the service?; What features of the service helped you to know where you were?; What strategy did you use for navigating through the service?; How did you feel using the service for the first time? Retrospective analysis (Nielsen, 1993) was integrated into the interview process as a means of facilitating participants' responses. This involved playing back any sections of the recorded interaction that were noted by the experimenter as being problematic. The participants listened to the recording and were asked to explain how they felt during the interaction, and to discuss any problems, hesitations, or incorrect navigation selections they had experienced. This method allowed participants to elaborate on preferences or issues that may have been raised by the usability questionnaire or the interview question, by referring to actual examples rather than hypothetical ones. The interview process was carried out in the following way: (i) A tape recording was taken of the participants' interactions whilst attempting the experimental tasks (ii) Simultaneously, the experimenter noted any problems experienced by the participant, with a corresponding tape counter number (iii) Participants were asked an open-ended question about some aspect of their interaction with the service, and were encouraged to respond as fully as possible (iv) Interaction events relevant to each interview question were then played back to participants to help spark memories, and enhance recall of how they felt at the time of the specific event.

7.4 Results

7.4.1 Participant variables

One-Way ANOVAs revealed no significant differences between experimental groups with respect to participants' age, previous mobile phone experience, previous fixed

line telephone experience, and previous computing experience. This suggests that the three groups did not differ in age, or technological background prior to the study. Table 7.3 shows descriptive data for participants' age, and previous telephone and computing experience.

Table 7.3 Descriptive data for the individual difference measures for experiment three

Participant variable	Minimum	Maximum	Mean	Standard deviation
Age	18.00	30.00	22.36	2.71
Mobile phone experience	9.71	58.00	34.90	11.73
Telephone experience	5.00	75.00	28.20	17.89
Computing experience	59.75	100.00	88.92	10.72

Table 7.4 shows the range of possible scores for the performance and attitude measures, as well as the grand mean and standard deviation for all users and across both services.

Table 7.4. Score range and means for the performance and attitude measures for experiment three

Measure Type	Measure	Range	Mean	SD
System performance measures	Time	0-100	55.82	9.88
	Task completion	0-100	88.78	16.56
	Prompt interrupts	0-100	92.61	9.79
	Nodes	>100	127.98	21.50
	Repeats	>0	0.43	0.67
Attitude measures	System response accuracy	1-7	5.13	0.65
	Likeability	1-7	4.78	0.68
	Cognitive demand	1-7	5.12	0.77
	Annoyance	1-7	3.75	0.97
	Habitability	1-7	4.83	0.96
	Speed	1-7	3.86	1.21

7.4.2 Performance measures

Data from the 5 performance measures were analysed using a 3 (experimental group) x 3 (task) mixed MANOVA to explore performance differences between the 3 experimental groups over the 3 tasks (see Table 7.5). The first 2 tasks were of a similar complexity, whilst the third was more complex involving backwards navigation through the service. The multivariate effect of task was marginally significant ($F(10, 26) = 2.13, p = 0.06, \text{Wilks' } \lambda = 0.55$) and significant univariate main effects for task occurred for all 5 user performance measures (Table 7.5), indicating that overall performance levels for the services tended to change over the

course of the 3 experimental trials. The multivariate effect of experimental group was not significant ($F(10, 62) = 1.45, p = 0.18, \text{Wilks' } \lambda = 0.66$). The multivariate effect for the interaction between experimental group and task was not significant ($F(20, 52) = 0.90, p = 0.59, \text{Wilks' } \lambda = 0.55$), suggesting that the manner in which performance changed over the tasks was largely consistent for all 3 services.

Table 7.5. Descriptive data for the performance measures for the 3 experimental tasks

Measure	Experimental group	Task 1		Task 2		Task 3		F value	Sig.
		Mean	SD	Mean	SD	SD	SD		
Time	Standard	58.97	8.78	51.03	15.99	55.26	9.07	7.94	<0-001
	Office	58.71	10.67	51.87	11.26	56.98	8.28		
	Windows	58.17	9.17	49.77	5.48	51.77	12.57		
Task completion	Standard	87.50	19.46	91.67	19.46	94.44	12.97	4.72	0.02
	Office	80.77	18.78	92.31	18.78	92.31	14.62		
	Windows	88.46	0.00	100.00	0.00	100.00	0.00		
Prompt interrupts	Standard	86.38	11.99	93.87	12.49	91.85	10.06	5.28	0.01
	Office	88.60	17.28	96.92	11.09	93.74	7.93		
	Windows	96.73	6.57	98.46	5.55	96.05	5.83		
Nodes	Standard	146.67	62.86	125.00	54.69	130.00	42.00	6.40	0.01
	Office	143.08	52.18	100.00	0.00	129.23	28.71		
	Windows	150.77	74.21	101.54	5.55	136.92	21.36		
Repeats	Standard	0.33	0.49	0.00	0.00	0.33	0.65	4.59	0.02
	Office	0.08	0.28	0.00	0.00	0.08	0.28		
	Windows	0.38	0.65	0.00	0.00	0.15	0.38		

7.4.3 Attitude measures

Data from the 6 attitude measures were analysed using a one-way independent groups MANOVA to explore differences in attitude between the 3 experimental groups. The multivariate effect of experimental group was not significant ($F(12, 68) = 0.88, p = 0.57, \text{Wilks' } \lambda = 0.75$). Simple univariate tests showed that only one of the six attitude measures was found to be significant, the measure likeability ($F(2,39) = 4.24; p = 0.02$). A post hoc Tukey HSD test showed that the computer desktop service was perceived as being significantly more likeable than the standard service ($p=0.02$), but there was no difference between the office filing system service and the standard service ($p=0.09$), or between the office filing system service and the computer desktop service ($p=0.60$). Although the effects were not significant, participants rated the computer desktop service most positively for 5 of the 6 subjective measures. The office filing system service was rated most positively for the subjective measure 'speed'. The standard service was rated most negatively for all 6 of the subjective measures (Table 7.6).

Table 7.6. Main effects between groups for the 6 attitude measures for experiment three

No	Subjective measure	Mean and SD	Standard service	OFS service	Windows service	F value	Sig.
1	System response	Mean	4.86	5.15	5.38	2.47	0.10
		SD	0.76	0.48	0.61		
2	Likeability	Mean	4.39	4.90	5.05	4.24	0.02
		SD	0.77	0.62	0.47		
3	Cognitive demand	Mean	4.74	5.24	5.40	3.10	0.06
		SD	0.93	0.63	0.61		
4	Annoyance	Mean	3.35	3.81	4.10	2.26	0.12
		SD	0.86	1.07	0.88		
5	Habitability	Mean	4.38	4.94	5.16	2.70	0.08
		SD	1.15	0.74	0.82		
6	Speed	Mean	3.70	4.02	3.86	0.24	0.79
		SD	1.16	1.32	1.24		

7.4.4 Recall of service structure

After completing the experimental tasks, participants were asked to recall the number of levels of menu options within the service, and the number of menu options at each of those service levels. This allowed comparisons to be made between services and was used in preference to a more complex test of service memory.

For the first of the memory measures, a one-way ANOVA revealed no difference between groups for the mean number of levels of menu options recalled by participants ($F(2,39)=0.15$; $p=0.86$) (standard: mean 4.14, SD 0.66; office filing system: mean 4.21, SD 0.80; computer desktop: mean 4.29, SD 0.61). Eight participants from each group correctly estimated that there were 4 levels of menu options.

The number of menu options at each level of the service was: level 1 = 2 options; level 2 = 2 options; level 3 = 3 options; and level 4 = 3 options. The sum of menu options within the 4 service levels was therefore 10 options. A one-way ANOVA was used to investigate whether any differences between groups existed for recall of the total number of service menu options. No differences in recall were found between groups ($F(2,39)=0.85$; $p=0.44$) (standard: mean 10.00, SD 1.66; office filing system: mean 9.64, SD 1.34; computer desktop: mean 10.36, SD 1.34).

7.4.5 Visualisation of the service structure

As part of the post task interview, participants were asked whether they had visualised the service structure and features whilst completing the tasks. Participants' responses to this question were cross-referenced with their interview responses regarding visualisation, in order to decide whether they had fulfilled the visualisation criteria derived from Ekstrom et al. (1976), and which were previously stated in chapter 6. Participants were then divided into those who visualised the services, and those who did not.

The ratio of visualisers to non-visualisers for each service was: standard service 3:11; office filing system service 11:3; and computer desktop service 12:2. A χ^2 test was performed on this data which suggested that significantly more participants in the metaphor-based groups visualised the service structure than in the standard group ($\chi^2(2, n=42) = 14.74, p < 0.001$). Cramer's V was found to be 0.59, suggesting that approximately 35% of the variation in frequencies of visualisers can be explained by the experimental group to which they were assigned. It can therefore be concluded that there was a significant association between experimental group and visualisation, with participants using the metaphor-based services significantly more likely to visualise the service structure than those using the standard service.

7.4.5.1 Performance measures

In order to examine differences in performance between visualisers and non-visualisers, a MANOVA was performed on the performance data pooled for all 3 services using visualisation as the dichotomous independent variable (Visualisers $n=26$, non-visualisers $n=16$). A significant multivariate effect of visualisation was found ($F(5, 36) = 3.17, p = 0.02, \text{Wilks' } \lambda = 0.69$). Simple univariate tests revealed that visualisers performed significantly better than non-visualisers on two of the performance measures: time, and prompt interrupts (Table 7.7).

Table 7.7. Main effects between visualisers and non-visualisers for the performance measures for experiment three

No	Objective measure	Mean and SD	Visualiser	Non-visualiser	F value	Sig.
1	Time	Mean	53.62	59.41	3.62	0.03
		SD	8.92	10.60		
2	Task completion	Mean	90.66	85.71	0.88	0.18
		SD	13.37	20.87		
3	Prompt interrupts	Mean	96.38	86.48	13.10	<0.001
		SD	7.29	10.43		
4	Nodes	Mean	125.96	131.25	0.59	0.22
		SD	18.87	25.53		
5	Repeats	Mean	0.35	0.56	1.04	0.16
		SD	0.63	0.73		

7.4.5.2 Attitude measures

In order to examine differences in attitude between visualisers and non-visualisers, a MANOVA was performed on the attitude data from all 3 services. The multivariate effect of visualisation was not significant ($F(6, 35) = 0.99, p = 0.45, \text{Wilks}' \lambda = 0.85$), although simple univariate tests revealed that for 3 of the attitude measures visualisers were significantly more positive towards the services than non-visualisers: system response accuracy, likeability, and cognitive demand. Although not a significant finding, visualisers expressed more positive perceptions towards the remaining 3 subjective measures than participants who did not visualise the service (Table 7.8).

Table 7.8. Main effects between visualisers and non-visualisers for attitude measures for experiment three

No	Subjective measure	Mean and SD	Visualiser	Non-visualiser	F value	Sig.
1	System response	Mean	5.30	4.86	4.94	0.02
		SD	0.44	0.84		
2	Likeability	Mean	4.93	4.53	3.72	0.03
		SD	0.59	0.76		
3	Cognitive demand	Mean	5.32	4.80	4.99	0.02
		SD	0.61	0.91		
4	Annoyance	Mean	3.87	3.54	1.20	0.14
		SD	0.98	0.95		
5	Habitability	Mean	5.01	4.53	2.63	0.06
		SD	0.76	1.18		
6	Speed	Mean	3.91	3.77	0.71	0.35
		SD	1.20	1.27		

7.4.6 Qualitative data

Theme-based content analysis (TBCA) was selected as an appropriate qualitative data analysis methodology to analyse the data resulting from the combined interviews and retrospective analyses. TBCA comprises a hybrid of qualitative and quantitative approaches that allows the grouping of data into meaningful themes, whilst also retaining the raw data throughout the analysis process, an important feature of qualitative analysis noted by Miles and Huberman (1994).

There were two main factors that influenced the decision to use TBCA for the current experiment. Firstly, the auditory metaphor-based services in this experiment were designed to facilitate the internal visualisation and navigation of a virtual information space. The TBCA method has been previously used to evaluate virtual environments in which users were verbally guided (Neale and Nichols, 2001), and therefore shares a number of similarities with the current task environment, which requires exploration and interaction within a mediated environment. Secondly, TBCA is a flexible method allowing the analysis of data produced by a number of different methods to be analysed. This supports the use of both the retrospective analysis data and the open-ended interview data that were gathered from this experiment.

TBCA involves a two-stage refinement process for identifying raw themes, and then higher-order themes. Raw themes emerging from the data are independently established, and are then tested against the themes proposed by at least one other researcher. This allows the reliability of the themes to be evaluated by calculating the number of times the researchers agree as a percentage of all possible themes proposed (Miles and Huberman, 1994). For this experiment, a second rater with no previous knowledge of the specific research domain was asked to establish raw data themes for the data from two of the seven interview questions. This was done informally, and the overall rate of agreement was 75%, which is above the accepted standard of 70% proposed by Miles and Huberman (1994). Following this, the number of user comments contributing to the raw themes were counted, to give an indication of their relative frequency.

Interview data and related retrospective analysis data were analysed from interview questions for each of the three service types, and categorised according to whether

participants visualised the service or not (Table 7.9 to 7.15). Each table shows the occurrence of raw data themes for each service type for both visualisers and non-visualisers with indicative quotes given by the users. It must be noted that some participants did not contribute comments to the themes, whilst other participants contributed comments to more than one theme, which is the reason that the total number of participants for each theme (e.g., $n=x$) is not equal to the number of participants in each experimental group.

For users of the metaphor-based services, imagining the service as described by the service dialogue was the most common way of visualising the office filing system service (Table 7.9, theme 1.1), whilst those who used the computer desktop service visualised what was described but tended to personalise their visualisation by combining features of a real world computer desktop filing system (theme 1.4). This effect may be attributed to domain experience, with the highly computer literate participants attempting to synchronise features of the phone service with features of their personal computer filing system. Some participants in both metaphor-based services visualised a completely different structure to the one presented (theme 1.2), which may have been related to a lack of domain experience, making it difficult for them to base their visualisation on the corresponding real world structure (theme 1.5). Some participants using the standard service visualised a hierarchy of links and nodes despite the absence of any spatial description (theme 1.2). There was also an attempt by some participants to simplify the metaphor-based services by changing the interface objects and making them the same at each level of the service (theme 1.3).

Table 7.9. Question 1: What aspects of the service did you visualise?

Visualiser	Standard	Office	Windows
Yes	<p>(Theme 1.2) Visualisation of a different structure (n=3): 'I visualised a tree structure with links and nodes'</p> <p>(Theme 1.3) Visualisation of different interface objects (n=1): 'I visualised the control options like the options in a software application, with the close button on the right hand side, and a drop down menu of options on the left'</p>	<p>(Theme 1.1) Visualisation of the service as it was described (n=7): 'I saw myself in an office creating images of the things at each level of the service, and I saw myself doing all the actions, and visualised everything but not in great detail'</p> <p>(Theme 1.2) Visualisation of a different structure (n=2): 'I imagined myself using a computer. I went to the control panel and clicked on different folders. There was also an internet browser with a Back button, and Home button'</p> <p>(Theme 1.3) Visualisation of different interface objects (n=1): 'I changed the things at each level into shelves, so instead of filing cabinets I visualised different shelves at the top, middle, and bottom'</p> <p>(Theme 1.4) Personalisation of the interface (n=1): 'When I worked in summer there were filing cabinets everywhere, with stickers on drawers, so for me it was really clear, and I customised my picture of the service to match my experience'</p> <p>(Theme 1.5) Domain experience (n=1): 'I don't use filings cabinet much, so I had to think really hard about how filing cabinets are set up, and then change my idea of how the service works'</p>	<p>(Theme 1.4) Personalisation of the interface (n=8): 'I pictured the service in the same way I organise folders on my desktop, but I was just picturing folders and sub folders and icons, not Windows or files, and I positioned them from left to right'</p> <p>(Theme 1.1) Visualisation of the service as it was described (n=2): 'Felt like I was speaking to a computer. I pictured the objects at each level, and I visualised the icons, and clicked on them, which took me to folders. I am familiar with this domain so it came naturally'</p> <p>(Theme 1.3) Visualisation of different interface objects (n=2): 'I pictured blank squares to represent the icons, and lined them up from left to right. Selecting options was like ticking boxes'</p> <p>(Theme 1.5) Domain experience (n=2): 'The service could be a problem for people who are not computer users, or who have limited domain knowledge'</p> <p>(Theme 1.2) Visualisation of a different structure (n=1): 'I saw a tree image. When the service said desktop I pictured the root of a tree, not a real tree but a tree diagram. So I imagined left and right, thinking in terms of left and right trees and branches and nodes all the way'</p>

For those participants who reported visualising the service, visualisation was most often reported as the primary navigational strategy (Table 7.10, theme 2.1). Visualisation was also mentioned as being a particularly valuable strategy for participants who were under pressure, had made a mistake, or were navigating backwards through the service (theme 2.3). Some visualisers combined visualisation with other strategies such as remembering menu options, and word spotting (theme 2.2). Non-visualisers simply attempted to remember the service menu options (theme 2.4), or alternatively, used a listening and word spotting strategy (theme 2.5).

Table 7.10. Question 2: What strategy did you use for navigating through the service?

Visualiser	Standard	Office	Windows
Yes	(Theme 2.1) Visualisation (n=2): 'I visualised the structure like a tree with links, which was helpful for navigation' (Theme 2.2) Mixed strategy (n=1): 'I visualised the service and remembered the numbers as directional, either up or down'	(Theme 2.1) Visualisation (n=6): 'You imagine yourself going into folders and drawers and looking for the information you need' (Theme 2.2) Mixed strategy (n=5): 'I was word spotting in conjunction with visualising the basic office structure'	(Theme 2.1) Visualisation (n=7): 'I visualised an overall picture of the service, and then used the same strategy as navigating on my computer' (Theme 2.3) Error recovery (n=4): 'The visualisation became more important especially when I made errors and had to go back'
No	(Theme 2.4) Remembering menu options (n=5): 'I remembered the sequence of options' (Theme 2.5) Word spotting (n=4): 'I was just listening out for the right words' (Theme 2.3) Error recovery (n=1): 'When I made a mistake or was under pressure, I attempted to visualise the service'	(Theme 2.4) Remembering menu options (n=2): 'I was just remembering a linear ordering of menu options, left, right, etc. The first level was easy to remember but after that it got more confusing' (Theme 2.5) Word spotting (n=1): 'I just listened to options and word spotted'	(Theme 2.4) Remembering menu options (n=2): 'There weren't many options to remember which made it easier' (Theme 2.5) Word spotting (n=1): 'I was just listening to the whole dialogue and word spotting'

Structural cues provided by the metaphor-based services were commonly reported by visualisers to help with orientation within the service (Table 7.11, theme 3.2). As the interface objects described at each level of these services were different, the overall structure of these services helped participants to work out where they were in relation to the start and end point of the services. Structural cues also appeared to help non-visualisers to orient themselves within the service.

Table 7.11. Question 3: What features of the service helped you to know where you were?

Visualiser	Standard	Office	Windows
Yes	(Theme 3.1) Control options (n=2): 'The Back function helped me to know where I was in the service'	(Theme 3.2) Logical service structure (n=9): 'Because it is an office structure it starts off at the filing cabinets, so you know you're at the beginning of the service, and then you are at a different place at each level' (Theme 3.1) Control options (n=2): 'You just go back and you know where you are' (Theme 3.4) Diagram of service (n=1): 'It would be good to give users a visual overview of the service structure to look at before using it'	(Theme 3.3) Service feedback (n=5): 'Whenever I chose an option the service told me where I was' (Theme 3.1) Control options (n=4): 'The Repeat and Back functions' (Theme 3.2) Logical service structure (n=3): 'Can get to learn the pattern because it is a logical progression through the service'
No	(Theme 3.3) Service feedback (n=5): 'The service tells you clearly where you are and which option you have chosen' (Theme 3.1) Control options (n=3): 'You always had option to go Back or Return' (Theme 3.6) Service complexity (n=1): 'It was clear enough to know where you were. I didn't feel I had to find my way out of a labyrinth by visualising it as the service was not complex enough to need a map or a visualisation'	(Theme 3.1) Control options (n=1): 'The extra menu options, because you could always go Back or Repeat' (Theme 3.2) Logical service structure (n=1): 'You were at a different place at each level'	(Theme 3.2) Logical service structure (n=1): 'The structure provided helped me to know where I was' (Theme 3.5) Limited options (n=1): 'There weren't many options to remember which made it easier'

Visualisers generally perceived themselves as having a good memory of the service structure by the time they had completed all of the tasks (Table 7.12, theme 4.1). Participants who visualised the services were less afraid to make mistakes (theme 4.2) than the non-visualisers, who had a tendency to listen to all of the menu options before offering a response rather than risk making a mistake (theme 4.5). This suggests that because visualisers perceived themselves to have a good memory of the service structure, they were more willing to explore the service and more confident about recovering from mistakes. Some of the non-visualisers simply did not bother to remember the service structure, as their strategy of word spotting did not require it (theme 4.4).

Table 7.12. Question 4: How well did you remember the structure of the service from task to task?

Visualiser	Standard	Office	Windows
Yes	(Theme 4.1) Good memory (n=4): 'I picked it up gradually, and by the end my memory of the structure was good'	(Theme 4.1) Good memory (n=7): 'It didn't take long to get a good idea of the structure' (Theme 4.2) Not afraid to make mistakes (n=3): 'I made mistakes but I learnt from that, and I knew I could recover easily from mistakes' (Theme 4.3) Poor memory (n=2): 'The different options at each level made it more complex, and I forgot which options went with which level'	(Theme 4.1) Good memory (n=12): 'I remembered the logical progression of objects, and can now remember most of the service structure. I transferred my GUI experience which helped'
No	(Theme 4.1) Good memory (n=2): 'I remembered the structure fairly well, but it improved with subsequent usage' (Theme 4.3) Poor memory (n=2): 'Not well. The structure stayed in short term memory, but then went again' (Theme 4.5) Listened to all options (n=2): 'I still listened to all the options because I didn't want to make a mistake'	(Theme 4.4) No attempt at memorising (n=2): 'I didn't bother to remember the service structure as I was just word spotting' (Theme 4.5) Listened to all options (n=2): 'I listened to all the options anyway just to make sure. I didn't want to make a mistake'	(Theme 4.1) Good memory (n=1): 'I remembered the basic sequence of the service'

Visualisers were more comfortable when using the services for the first time than were the non-visualisers, who generally felt uncertain (Table 7.13, theme 5.2). Visualisers stressed the importance of previous domain experience as a factor affecting first time usage (theme 5.3) suggesting domain experience to impact visualisation ability. Non-visualisers were more likely to perceive the services as cognitively demanding on first time use (theme 5.4), suggesting visualisation to be potentially less demanding on cognitive resources.

Table 7.13. Question 5: How did you feel using the service for the first time?

Visualiser	Standard	Office	Windows
Yes	(Theme 5.1) Uncertain (n=2): 'I was a little confused to start with' (Theme 5.3) Domain experience (n=1): 'I know about these kind of services so I felt fine, some people may not like talking to a computer though'	(Theme 5.2) Comfortable (n=6): 'Liked it right from the start, and I was interested because it's a new idea' (Theme 5.3) Domain experience (n=4): 'If you are familiar with these services in general, you would feel very comfortable'	(Theme 5.3) Domain experience (n=3): 'If you don't use these services or are not a regular computer user it would be a bit confusing' (Theme 5.1) Uncertain (n=2): 'The first time I was a bit uncertain because I don't know what menus are' (Theme 5.2) Comfortable (n=2): 'The first time I used it I thought it was friendly and straightforward to use'
No	(Theme 5.1) Uncertain (n=7): 'I felt a bit anxious the first time' (Theme 5.2) Comfortable (n=2): 'I felt fine after hearing the menu options for the first time' (Theme 5.4) Cognitive demand (n=2): 'I felt irritated because too much concentration was required'	(Theme 5.1) Uncertain (n=2): 'I felt overwhelmed, seemed strange, I didn't understand the concept' (Theme 5.4) Cognitive demand (n=1): 'The quantity of dialogue made it distracting to start with, and harder to remember the options'	(Theme 5.1) Uncertain (n=1): 'I felt a little dubious to start with' (Theme 5.2) Comfortable (n=1): 'I was happy. It was easy and simple to use and pick up for the first time'

Although question 6 was intended as an open-ended question, it became a fixed response question, with participants simply stating the task number at which they became confident, and not expanding on their answers. The answers in Table 7.14 reflect the fixed nature of the responses. Participants who visualised the services generally felt confident using the services earlier in the experiment, typically after attempting the practice task (Table 7.14, theme 6.1). In contrast, the non-visualisers took longer to feel confident, typically after the first or second task (themes 6.2 and 6.3).

Table 7.14. Question 6: At what point did you start to feel confident using the service?

Visualiser	Standard	Office	Windows
Yes	(Theme 6.1) Early (n=3)	(Theme 6.1) Early (n=5) (Theme 6.2) Intermediate (n=4) (Theme 6.3) Late (n=2)	(Theme 6.1) Early (n=5) (Theme 6.2) Intermediate (n=4) (Theme 6.3) Late (n=1)
No	(Theme 6.3) Late (n=6) (Theme 6.2) Intermediate (n=3) (Theme 6.1) Early (n=1)	(Theme 6.1) Early (n=1) (Theme 6.2) Intermediate (n=1) (Theme 6.3) Late (n=1)	(Theme 6.2) Intermediate (n=2)

Participants who visualised the metaphor-based services preferred these services to the standard service, citing the main reason as being improved orientation (Table 7.15, theme 7.1). Other reasons given were that the metaphor-based services were faster (theme 7.5), more interesting (theme 7.6), and based on a familiar domain (theme 7.3). Some participants who visualised the office filing system service suggested that it might be more suitable for a more complex and extensive service than the current service, which may reflect the perceived usability benefits of visualising the service (theme 7.4). However, on the issue of complexity, a Windows service participant anticipated problems attempting to visualise a larger service with more menu options (theme 7.4). Non-visualisers generally showed a preference for the standard service because they were more used to it (theme 7.2).

Table 7.15. Question 7: How does this service compare to the standard menu style service?

Visualiser	Standard	Office	Windows
Yes	N/A	(Theme 7.1) Orientation easier with metaphor service (n=3): 'I prefer this service because it is less confusing and you know where you are' (Theme 7.5) Speed (n=2): 'I prefer this service because it is faster' (Theme 7.2) More familiar with standard service (n=2): 'I prefer the standard service because I am more used to it' (Theme 7.3) No difference (n=2): 'It wouldn't make any difference to me' (Theme 7.4) Service complexity (n=2): 'This service is better for getting more complex information from a more complex service. It seems over complicated for this service' (Theme 7.6) Interesting (n=1): 'I prefer this service because it is more interesting'	(Theme 7.1) Orientation easier with metaphor service (n=4): 'This is better because you know where you are, and it is easier to navigate' (Theme 7.7) Domain knowledge (n=3): 'I prefer this service because it is based on a familiar domain' (Theme 7.8) Dialogue length (n=2): 'I prefer the standard service because it has shorter dialogues' (Theme 7.4) Service complexity (n=1): 'I prefer this service, but if it was bigger with more menu options, it would be difficult to visualise lots of icons or windows'
No	N/A	(Theme 7.2) More familiar with standard service (n=2): 'I'm used to menu options as numbers, so I prefer the standard style of service' (Theme 7.3) No difference (n=1): 'If you are not visualising the service then they are both the same'	(Theme 7.1) Orientation easier with metaphor service (n=1): 'I prefer this service because you know where you are' (Theme 7.2) More familiar with standard service (n=1): 'I prefer the number service because I have more experience using it'

7.4.7 Design factors

Based on individual interview responses, an analysis of interview content was performed whereby individual comments were examined for underlying themes, meanings, and issues. This analysis revealed a number of themes relevant to the design of automated mobile phone services, and which have been listed in Table 7.16 as potential design factors. Factors 1-13 are also relevant to GUI design, but factors 14-18 are specifically relevant to the design of speech-based automated mobile phone services, and are discussed in more detail in the discussion section that follows.

Table 7.16. Design factors for automated mobile phone services

No	Factor	Definition
1	Error recovery	The degree to which features provided by the service enable a user to successfully recover from errors
2	Navigation	The degree to which navigational cues provided by the service prevent a user from getting lost, help a user to orient themselves, and to find the information they are looking for
3	Feedback	The degree to which information provided by the service adequately informs a user about the actions they have taken, and the results of those actions
4	Speed of service	A user's perceptions of the length of the service dialogue, and of how quickly the dialogue and menu options are generated
5	Previous experience	The degree to which previous experience in other real world or computer-based domains affects a user's perceptions and interactions with the service
6	Service consistency	The degree to which performing similar tasks involves similar actions and operations, and produces similar service output
7	Predictability	The degree to which a user's knowledge of the service gained from previous interactions can help them to predict the results of future interactions
8	Learnability	The degree to which a user perceives the service as being easy to learn to use
9	Memorability	The degree to which the service is easy to remember how to use once learned
10	Ease of use	The degree to which a user perceives the service as being easy to use
11	Service complexity	The degree to which a user's perceptions of the service complexity match their expectations of how complex the service should be for a particular domain
12	Cognitive demand	Refers to a user's perceived amount of cognitive effort needed to interact with the service
13	Customisability	Refers to either user-initiated (adaptability) or service-initiated (adaptivity) modifiability of the service interface
14	Repetitiveness	The degree to which a user perceives the service dialogue as being too repetitive, and not varied and interesting enough
15	Interaction strategy	Refers to the strategies adopted by users in order to successfully interact with the service, and to complete the desired tasks
16	Metaphor integration	Refers to the introduction of a new metaphor by a user to help them interact with some aspect of the service functionality
17	Likeability of service voice	The degree to which users perceive the service voice as being pleasant and friendly
18	Social acceptability	The degree to which a user feels comfortable using the service in public places

Table 7.17 contains interview responses from each of the 3 services relating to the 18 design factors, and which identify the key perspectives related to each service and highlights their relevance across the three services.

Table 7.17. Interview quotes from the 3 services supporting the 18 design factors

No	Factor	Standard	Office	Windows
1	Error recovery	'If you make a mistake you can go back or return, which makes it easy to recover'	'The help option should be offered earlier in the dialogue so you know it's there and don't panic'	'I just used the back function if I was lost and I immediately knew where I was'
2	Navigation	'The structure was like a tree, which was helpful for navigation'	'Because it is an office structure it starts off at the filing cabinets, so you know you're at the beginning of the service, and then you are at a different place at each level'	'You always know where you are with this service which makes it easier to navigate'
3	Feedback	'The service tells you clearly where you are and which option you have chosen'	'It's simple, and it tells you what you are doing, so it's hard to get confused'	'Whenever I chose an option the service told me where I was'
4	Speed of service	'It was too slow and too repetitive'	'It was too long and too formal. Also if you make a mistake it takes even longer to get to the information'	'The service was too slow, and she repeated herself too much which can get annoying'
5	Previous experience	'I know about these kinds of services so I felt fine. Some people may not like talking to a computer though'	'When I worked in the summer there were filing cabinets everywhere, with stickers on drawers, so for me it was really clear and I customised my picture of the service'	'I am familiar with this domain so for me it came naturally, but it could be a problem for novice users or people with limited domain knowledge'
6	Service consistency	'You always have the option to go back or return to the start'	'The service is logical, and I liked it because it was simple, using simple words'	'The menu options were the same at each level'
7	Predictability	'It's easy to go back, and easy to predict the options'	'Because the system described the structure so well, you could predict the options even though they were different at each level'	'You could jump ahead and say the options without listening to the whole dialogue'
8	Learnability	'I felt a bit anxious the first time, but after I had heard the voice and the options I felt fine'	'The concept of left/right menu options might take a while to pick up, but if you're used to using filing cabinets it would be no problem'	'I'm familiar with PC interfaces so it was easy and simple for me to learn'

9	Memorability	'By the last task I was OK, but I still listened to all the options because I didn't want to make a mistake'	'It was easier to remember the menu options by visualising them. Didn't take long to get a good idea of the structure'	'I remembered it pretty well because I arranged it similar to my own computer'
10	Ease of use	'It got clearer and easier as I went through the service and completed several tasks'	'It was easy to learn quickly, easy to follow, and the structure was clear'	'The service was really easy to use and consistent throughout'
11	Service complexity	'The service was not complex enough to need a map or visualisation. I didn't feel I had to find my way out of a labyrinth by visualising it'	'The different options at each level made it more complex, and because I wanted to predict options this made it frustrating and more difficult'	'The design is OK for a limited service, such as in a hospital or school, but for a bigger public service it would be a problem to extend it. You could maybe provide a diagram of the structure'
12	Cognitive demand	'It was irritating because I had to concentrate so hard on the service'	'I don't use filing cabinets much, so I had to think really hard about how they are set up, and then change my idea of how the service works'	'The visualisation aspect became more important especially when I was under pressure or had made errors and had to go back'
13	Customisability	'It's not flexible enough, and you can't deviate from the fixed options'	'If you were a more advanced user it would be good to have short cuts'	'The service should log the number of times you've visited and start to cut out some of the dialogue for advanced users, or offer short cuts'
14	Repetitiveness	'It's OK if you get it right the first time, but if you make a mistake and have to start again it starts to sound very repetitive'	'The service was too repetitive'	'She repeated herself too much which can get annoying'
15	Interaction strategy	'I remembered the order of menu options, but listened to the options all the way through just to make sure I got them right'	'I visualised different things at each level but not in much detail, just in general for example filing cabinets and drawers'	'I used the same strategy as I'd use for navigating on my computer'
16	Metaphor integration	'I saw the control options like the options in an application, with the close button on the right hand side and the other options in a drop down menu on the left hand side'	'I imagined going to the control panel like on a computer, then clicking on a folder. I also visualised a browser with a back button, and a home button'	'I thought about it mainly in terms of folders – so folders within folders. I pictured blank squares to represent the icons lined up from left to right. Selecting options was like ticking boxes'
17	Likeability of service voice	'The voice was irritating and makes the service sound unfriendly'	'A more relaxing voice would have been better. The voice was too computerised'	'The voice was boring and gets irritating after a while. This may affect older users more'

18	Social acceptability	‘I would feel embarrassed to use the service in a public place’	‘I’d feel fine because it would be obvious I was using an automated service. Some people may feel a bit embarrassed’	‘If I used it on a train or bus I would feel uncomfortable and a little silly’
----	----------------------	---	--	--

7.5 Discussion

No overall performance or attitude differences were found between the three services suggesting that the use of metaphors is not necessarily advantageous on first use. This finding is consistent with the results from experiment two, which showed no subjective preferences for metaphor-based services, and significant improvements in user performance only to occur after prolonged periods of use. Data from this experiment also suggest that phone services based on a computer graphical user interface fare no worse than current numbered phone services on first use when considering user preference and performance, which may indicate that such metaphors could be successfully applied to voice interfaces without any performance detriment to users. However, it must be noted that the participant sample were all computer science students with much higher than average computer literacy, and greater familiarity with computer related terminology. They may then have preferred the computer desktop service due to their initial familiarity with the terminology, and their high levels of experience with the corresponding GUI version of the interface, which may have led them to be more positive about the service, and to perform better with it relative to the other services. This finding does not undermine the benefits of using a HCD process, which was the process used to select and develop the office filing system service, for the following reasons. Firstly, the computer desktop metaphor was a metaphor that was well known by all of the participants, due to their Computer Science backgrounds, and was therefore a metaphor that was likely to have resulted in high acceptance rates. Secondly, the HCD process that had been used to design and define the office filing system service dialogue, was used as a template for the computer desktop service, without which the appropriate dialogue flow, structure, and length would have been unknown. Yankelovich et al. (1995) found that the vocabulary used to describe a GUI does not transfer well to a speech-based interface, but the findings of this experiment suggest that for highly computer literate users, this may not be strictly true.

The computer desktop service showed the best performance levels on 3 measures (time, task completion, prompt interrupts) and those using the office filing system service showed the best performance levels on 2 measures (nodes, repeat). The computer desktop service also showed the poorest performance for the number of nodes used. Whilst successfully completing the highest number of tasks, participants using the computer desktop metaphor were more likely to interrupt prompts, thereby taking less time, but in doing so making more mistakes and consequently leading them to use more nodes. In contrast, participants using the office filing system tended to listen to the service dialogue for longer, used the 'repeat' function very rarely, and used the fewest number of nodes to successfully complete tasks. The standard service participants exhibited the poorest performance, and can be characterised by their tendency to listen to more of the service dialogue before interrupting, taking more time, and requesting the highest number of dialogue repeats with the lowest levels of successful task completion. Different styles of interaction were therefore evident between services.

No statistical differences in performance were found between groups, or over the three tasks, suggesting that none of the services offered any significant performance benefits when tackling simple or more complex tasks. According to Mayer (1981) and Borgman (1986), complete and accurate mental models play a relatively important role in complex task performance, while having less impact for simple task performance. If the metaphor-based services had allowed participants to develop more complete mental models of the services, then it would be expected that performance would be improved for the complex task. A possible reason for why this effect was not evident in the current experiment was that users did not find the service complex, or that differences in the difficulty levels between tasks was not sufficient to highlight the potential performance benefits of metaphor. Evidence from the interviews suggests that, especially for the office filing system service, some participants felt the metaphor-based style of service design would have been of more benefit if used to implement a more complex and extensive service.

Participants perceived the standard service least positively on all attitude measures, and perceived the computer desktop service most positively on all attitude measures, apart from speed. On this measure the office filing system service was perceived as

being the fastest. A potential explanation for this might be made by reference to the real-world source domains of the metaphor-based services. When using a desktop GUI on a computer, if the user knows where the relevant menu item is located within the application interface, a task can be performed quickly by moving the cursor and clicking on the item. However, when this GUI interface is extended to a speech user interface, the process of navigating to, and selecting the desired menu item may be slowed down by the sequential nature of the message prompts, compared with the real world metaphor referent. With reference to the office filing system metaphor, the speed with which a user might physically locate a folder or file from within a real filing cabinet would be comparatively slow. Actions such as opening a drawer, and flicking through partitions to find information are inherently more time consuming than pointing and clicking with a cursor. Therefore, participants may have perceived the office filing system service to be fast compared to its real world equivalent.

The metaphor-based services were visualised by significantly more participants than the standard service, which supports the use of spatial metaphor as an aid to conceptualising a system in the form of a visualisation. Due to the high number of visualisers in both metaphor-based services, and the low number in the standard service, it may be argued that even those participants with relatively lower visualisation ability were able to visualise the metaphor services, whilst only those with very high visualisation ability attempted to visualise the standard service. Across the three services, participants who visualised the service structure performed significantly better than those who did not. This result suggests that visualisation of a phone service does indeed aid the cognition of the service by promoting the development an internal mental model of the service, which can be used to guide interaction, and that such visualisation is more likely to occur in a metaphor-based system.

The qualitative reports show that the majority of participants visualised what was described by the service but with limited attention to direct detail. Previous experience or knowledge of the real world equivalent of the metaphor domain had a major effect on participants' ability to visualise, and on the content of that visualisation. This finding is in accordance with the theory of interface metaphor usage (Carroll and

Mack, 1985), which relies on a prior understanding of the metaphor domain for knowledge transfer to occur.

Participants using the computer desktop service were most likely to personalise the service, adapting and enhancing their visualisations using features derived from their real world experiences, which may be explained by their high levels of computer literacy. In some cases, additional features (or images) were integrated into the service metaphor to help participants conceptualise certain aspects of the service structure or functionality. Data from the computer desktop service also emphasised a tendency for some participants to spatially arrange non-spatially presented information. Objects within this metaphor were not presented with spatial orientations, but participants who visualised the service often reported arranging the interface objects spatially in a similar way to how their own computer desktops were arranged. This suggests that some participants will tend to think in a visual way about a domain that is not only non-visual, but is also presented non-spatially. For such participants it might be expected that the benefits of spatial metaphor will be most powerful.

Those participants using the office filing system service were most likely to visualise what was described by the service. Despite their low levels of actual filing cabinet experience, these participants had a good knowledge of how they work, which may be a benefit of this metaphor compared to the computer desktop metaphor which requires knowledge of both the domain and the terminology. Some participants using the standard service visualised the service as a hierarchy of nodes and links, even though the interface provided no spatial or navigational cues. This may highlight a tendency for some participants towards visualisation of speech-based services, even in the absence of any aids to visualisation.

Interviews revealed visualisation to provide a number of usability benefits. Firstly, it aided navigation of the services, especially when participants had made a mistake and felt under pressure. Secondly, when using services on their first attempt, visualisers generally felt more comfortable. Moreover, visualisers felt more confident about their ability to use the services after fewer exposures to the service. Thirdly, visualisation helped participants to orient themselves by providing structural cues. Of the participants who visualised the metaphor-based services, the most commonly cited

reason for preference of these services compared to the standard service was orientation, which gave participants an improved perception of their memory of the service structure and appeared to promote exploration by diminishing concerns about making mistakes, thus increasing overall confidence.

Due to the high number of visualisers across groups, specifically in the metaphor-based services, it can be concluded that participants were able to visualise the services even though in a public location. When using a mobile phone in public, a user will often be subject to distraction and competing cognitive activity. It is at these times that a user requires a system that is less cognitively demanding to use. Interview data from some participants suggests that when these participants perceived themselves as being under pressure, such as when they were in a busy shopping mall with high levels of visual and auditory distraction, they were more likely to visualise the service. This lends support to the idea that a visualisation can offload cognitive mental processes to perceptual processes, thereby expanding working memory capacity. The internal visualisation of a spatial mental model therefore appears to be compatible with real world vision, even in highly distracting environments.

The secondary analysis of the qualitative data revealed the presence of 18 factors that may be important to consider when designing speech-based mobile phone services. Of these, the five factors that were judged to be of particular relevance to speech-based phone services were repetitiveness, interaction strategy, metaphor integration, likeability of the service voice, and social acceptability. In terms of repetitiveness, participants were intolerant of service dialogue that was too repetitive, in the same way that they would be intolerant of a real person that kept repeating himself or herself. Implications for design would therefore be to reduce the quantity of dialogue for experienced users, and to offer a reduced version of the dialogue when the 'Repeat' function is accessed, thereby introducing an element of adaptability and variety into the dialogue.

Interaction strategy refers to the strategy adopted by the participant in order to complete the task. The data suggests that three main strategies were used: word spotting, visualisation of the service structure, and memorising the menu options. These strategies were either used in isolation, or were sometimes combined, for

instance, a participant may have used a combination of visualisation and memorisation of menu options. When designing such services, it may therefore be beneficial to consider the various ways that a user will approach navigation of the service, thereby increasing the usability of the services for the full range of potential users.

Metaphor integration refers to the extent to which a participant integrates a new metaphor with the existing metaphor to help them with some aspect of the service functionality. This process appears to be more salient to non-visual speech-based systems. In metaphor-based GUIs, the metaphor is concrete and visual, and it may therefore be less likely that the user will deviate from, or adapt the metaphor presented. However, with a metaphor-based speech service, the dialogue suggests a metaphor that can be visualised by users. The user therefore controls the exact nature of the visualisation, and has the freedom to adapt the metaphor that is visualised, or to integrate new features. It is therefore necessary for the designer to be aware of the ways in which a user will attempt to extend and adapt the metaphor provided, in order to provide optimum support for the service functionality.

Likeability of the service voice was a feature of the services that was consistently raised by participants in the post-task interviews, and in general participants both disliked the synthetic voice and found it difficult to comprehend. These points became less important after repeated usage, as participants found that they could ‘tune into’ the voice, but for first time usage likeability of the voice could be a critical factor affecting service acceptance. Pilot studies using different voices would therefore help to determine the type of voice preferred by users.

Social acceptability refers to whether the participant felt comfortable using the service in public. This may be related to their overall attitude towards mobile phone usage in public, or more specifically, related to the wording of the menu options that they are required to say in order to progress through the service. Due to the public nature of speech, the wording of the menu options that a user is required to say appears to be a critical factor influencing users’ levels of public comfort when using the services. Although mobile phone services are not always accessed from public places,

consideration of the acceptability of the menu options used in services may therefore be advisable.

Three main implications for designers of speech-based mobile phone services can be derived from this experiment. Firstly a metaphor that organises information spatially will be expected to lead to usability benefits for the design of a speech-based mobile phone service. Secondly, since even those who have a low ability to visualise services seemed to benefit from the introduction of service metaphors, steps could be taken within system design to support users who have difficulty constructing visual system images, possibly through the use of a graphical overview of the service structure. An alternative approach would be to enable participants who do not like the metaphor, or those who find it difficult to visualise, to turn the metaphor off and simply use a number-based version of the service. Thirdly, through the use of spatial metaphor it may be possible to offer services that have a more complex menu hierarchy. Bond and Camack (1999) suggest that telephone voice menu systems should be based on a hierarchy of no more than three levels deep, and that no more than four menu options should be presented at each point within the service. Whilst four menu options at any given level may be appropriate, findings from the present experiment suggest that users may be able to cope with more than three levels within a hierarchical structure particularly when using the qualitatively different visual images and properties invoked by a relevant layered spatial metaphor. Such findings may be of relevance to more complex phone systems (e.g. telephone banking), which require a larger number of levels to complete transactions and provide a comprehensive customer service.

7.6 Chapter summary

Users who visualised the phone services performed significantly better than those who did not, or were not able to visualise the phone services on first use. The use of spatial metaphors within a hierarchically structured mobile phone service promotes visualisation, where spatial metaphors are appropriately designed and relevant to service content. The subsequent visualisation may be the factor that led to improved attitudes towards and performance with the metaphor services relative to the standard number-based service. Metaphors borrowed from a computer graphical user interface may be successfully transferred to speech-based phone services without any obvious detriment to user performance on first use for the experienced computer users who

took part in this experiment. However, metaphors should be relevant to participants. Whereas participants with no previous experience of using a filing cabinet might be able to understand the service, participants with no previous computing experience might experience problems with the terminology and structure of a computer-based metaphor. The challenge for designers of speech-based mobile phone services is to provide an appropriate spatial organisational metaphor containing navigational cues that allows visualisation of the service, but does not lead to overlong dialogues, and which does not obstruct the information the user is trying to access.

:: CHAPTER 8

Conclusions

8.1 Summary of the experimental findings

This thesis set out to investigate whether interface metaphor could be used to improve the usability of speech-based automated mobile phone services. The widely accepted four-component model of HCI, proposed by Eason (1991), was used to frame the investigation. An investigation of the interface design of automated mobile phone services addressed the ‘computer’ component of the model, and was driven by the research question: Can interface metaphors improve the usability of speech-activated automated mobile phone services? Measurement of a range of users’ individual differences addressed the ‘human’ component of the model, and sought to answer the research question: To what extent do the individual characteristics of users affect the usability of metaphor-based speech-activated automated mobile phone services? An investigation of the physical and social contexts of use in which such services are used addressed the ‘environment’ component of the model, and was designed to answer the research question: Does context of use affect the usability of metaphor-based speech-activated automated mobile phone services? The final ‘task’ component of the model refers to the tasks that users were required to perform with the services, the data from which enabled the other three components of the model to be

investigated. The following three sections summarise the key findings of this thesis relating to three components of the model: computer, human, and the environment. The implications of the findings are then considered with respect to relevant theory, and located within the context of previous studies conducted within the field of HCI.

8.1.1 Computer component: Interface metaphor

The first stage of the experimental work conducted for this thesis was to generate a number of potential metaphors for automated mobile phone services, and then to select and develop a sub-set of these metaphors for a mobile city guide service. In the Dutton et al. (1999) study, which was the only previous study to have investigated the use of interface metaphors for telephone-based interfaces, the metaphors were experimenter generated, and based on the experimenter's intuition. Preliminary study one of this thesis used a human-centred approach to generate, select, and develop metaphors.

This approach revealed that card sorting was a usable and productive technique that may be recommended as an important visual means of stimulating metaphorical thought about telephone-based interfaces, and thus facilitating the selection of appropriate metaphors. As part of their methodology for developing interface metaphors, Neale and Carroll (1997) recommended visual techniques, such as sketching, for developing metaphors. Their approach is supported by the results from preliminary study one, which revealed that card sorting rather than sketching may be a more effective visual technique.

To develop the features of a selected metaphor, Alty et al. (2000) proposed an unstructured technique they refer to as 'fruitful conversation'. To provide a more rigorous and structured method for generating relevant language to describe a metaphor, the POPITS model (Cates, 2002) was used within preliminary study one of this thesis. The advantage of the POPITS model is that it decomposes a metaphor into six key areas of attributes, providing a more thorough approach than simply considering the metaphor as a single entity. Card sorting and the POPITS model were found to be complementary techniques, with the card sorts produced by participants providing a reference point that could be used to both stimulate, and guide, the generation of metaphor features within the POPITS model. A significant number of

the metaphor features generated in this way were incorporated into participants metaphor-based task explanations. This provides evidence of the salience of the features generated to the design of a metaphor-based system, and supports the use of these techniques as an addition to step two of the Alty et al. (2000) framework. There exists no previously documented use of the POPITS model within a metaphor design process. The work reported for this thesis therefore provides empirical support for the technique, and demonstrates that it may be used for speech interface development, as well as graphical user interface development.

Another finding from preliminary study one was a coherent categorisation of interface metaphors for speech-based automated mobile phone services. Previous research has focussed on the formulation of metaphor categories for graphical user interfaces, such as the three categories proposed by Condon and Keuneke (1994): spatial, activity-based, and interactional. The five metaphor categories resulting from the work conducted for preliminary study one of this thesis formed a sub-set of the spatial category formulated by Condon and Keuneke (1994), and were: Hierarchical, Shopping venue, Transport system, Information provider, and Natural circular. Descriptions of the features of the categories can be seen in section 4.2.3.6. These categories may provide designers with a useful basis for considering which metaphors to use for the design of speech-based mobile phone services, by limiting the range of appropriate overall structures and individual metaphors.

In experiment one, three metaphors from the three different metaphor categories that were rated by participants as being most applicable to mobile phone services, were selected and used to implement three metaphor-based versions of a mobile city guide service. The usability of the metaphor-based services was then compared to that of a standard number-based service, which was designed in the same style as commercially available automated phone services. It was found that participants both preferred, and performed better with, the office filing system service. This effect was not evident on first time usage, suggesting that participants may require a number of exposures to learn and to accept a metaphor-based interface that has replaced a non-metaphor interface for a known system. Participants already had experience of using standard style services, and therefore had to relearn how to use a familiar service using an unfamiliar metaphor-based interface. Therefore, in experiment one, the

impact of metaphor for learning a new system was not being evaluated. Rather, the impact of metaphor for learning, and improving performance with a known system was being evaluated. Despite this, for the metaphor-based interfaces to emerge with higher levels of usability after three exposures is an indication of the power of metaphor as a learning aid, and as a technique for improving usability.

In experiment two, the use of the office filing system metaphor-based service was compared to the non-metaphor standard service over a longer period of time, to assess the effect of metaphor on the retention of the services. The office filing system service significantly improved overall performance relative to the standard service, whilst performance for first time usage also tended to be better. This suggests that the metaphor was more effective at helping users to retain knowledge about the service over time. Approximately two thirds of participants reported visualising the office filing system service structure. Visualisation leads to the formation of a mental model (Ware, 2000). It may therefore have been the visualisation of a mental model that led to performance improvements with the metaphor-based service. Visualisation also offers a means of enhancing cognition (Larkin and Simon, 1987; Norman, 1993; Card et al., 1999), which may also help to explain the performance gains shown by the metaphor-based service in experiment two. The nature of visualisation was subsequently explored in experiment three.

The results from experiment three revealed that the metaphor-based services were visualised by significantly more participants than the standard service, suggesting that metaphor-based services lend themselves to easier visualisation. Moreover, participants who visualised both the metaphor-based services and the standard service performed significantly better than those who did not. Participants reported that visualisation aided navigation, made them feel more comfortable and confident on their first exposure to the services, and helped them to orient themselves by providing structural cues. Visualisation was also the most commonly offered reason for a participant's preference for a metaphor-based service. The visualisation strategy therefore enabled users to develop a mental model of the service consistent with the designer's model (Norman, 1986), which helped users to both learn the service, and to retain that knowledge over time.

Experiment three also examined the suitability of a computer desktop GUI metaphor, adapted from the computing domain, for a speech-based mobile city guide service. The metaphor was not developed using a HCD process, but was found to successfully transfer to the mobile city guide service without any attitude or performance detriment, suggesting that computing metaphors may have potential for the design of speech-based applications. However, rather than undermining the value of the HCD process that was used to generate and develop the office filing system metaphor, this result highlights the fact that, for a metaphor to be effective, users must be familiar with the metaphor. In this case, users were participants from a Computer Science department, and used the graphical equivalent of the desktop metaphor regularly, which enabled them to successfully transfer the metaphor to the new system. However, the fact that the desktop metaphor may not be applicable to all users is highlighted by the low ratings it was awarded in preliminary study one. The desktop metaphor service also benefited from the human-centred dialogue design process used to develop the office filing system service, which provided a dialogue template for the design of the dialogue for the desktop metaphor service. This further reinforces the value of a HCD process.

A further outcome from experiment three was the identification of five factors relevant to the design of speech-based mobile phone services, and which may offer designers additional guidance on design. The five factors were repetitiveness, interaction strategy, metaphor integration, likeability of the service voice, and social acceptability. Definitions of these factors, and the implications they raise for designers were discussed in section 7.5. The ‘repetitiveness’ factor is related to the ‘annoyance’ factor proposed by Hone and Graham (2000), whilst the ‘likeability of the service voice’ factor is related to the ‘likeability’ factor proposed by Hone and Graham (2000). Despite this, these two factors may offer guidance on the relative importance of sub-factors within the six main factors suggested by Hone and Graham (2000). The other three factors generated are distinct from the six Hone and Graham (2000) factors, and it is possible that they have the potential to extend the six-factor Hone and Graham (2000) model, specifically for the evaluation of metaphor-based speech systems.

Despite the widespread use of interface metaphor for interactive systems, there have been few empirical studies that have found significant performance benefits of a metaphor-based system relative to a non-metaphor system. The work conducted for this thesis addresses this dearth of evidence in a number of ways. Significant performance benefits were found for a mobile phone service based on an office filing system metaphor relative to a non-metaphor version of the same service. This finding is consistent with one previous study, conducted by Dutton et al. (1999), which analysed the effects of metaphor on the usability of an automated telephone home shopping service. They found performance advantages for the metaphor-based versions of the service relative to the non-metaphor service. The work conducted for this thesis therefore extends the usability benefits of metaphor from automated fixed line telephone services requiring keypad input, to automated mobile phone services requiring speech input. Moreover, within more general empirical studies of metaphor, the work reported in this thesis supports the results of a study conducted by Smilowitz (1996). Smilowitz (1996) found performance benefits for a version of a web browser based on a metaphor, compared to a version that was literal and free from metaphorical associations. Although the office filing system metaphor service developed for this thesis resulted in performance benefits, it must be noted that different versions of the office filing system metaphor service were not compared. Alternative implementations based around the same metaphor may therefore have further improved the implementation tested.

The experimental work conducted for this thesis revealed two important implications for designers of speech-based automated mobile phone services. Firstly, that choosing an appropriate interface metaphor that organises information spatially, and that was generated as part of a HCD process, will be expected to improve performance with the service by enabling users to visualise a mental model of the service structure. Secondly, that choosing an appropriate spatial interface metaphor will allow services to be designed that have a deeper, more complex menu hierarchy than the current recommendation of three levels (Bond and Camack, 1999), by providing navigation and orientation cues.

8.1.2 Human component: Individual differences

An investigation of the effects of participants' individual differences on service usability was conducted as part of experiment two. There were found to be no significant predictors of attitude or performance with the metaphor-based office filing system service. Gender, working memory, and previous telephone experience were found to be significant predictors of attitude towards the standard service, whilst gender was found to be a significant predictor of performance with the standard service when used in public locations but not when used in private locations.

Female participants generally perceived the service more positively than male participants, but in public, despite their more negative perceptions of the service, male participants performed better than females. A possible explanation for this effect may be related to males' perceptions towards the usefulness of the service. Male participants perform better with systems that they perceive as being useful, whereas ease of use is a better predictor of performance for female participants (Venkatesh and Morris, 2000). Male participants may therefore have perceived the service as being useful, but not as being usable, according to the subjective factors measured for the work conducted for this thesis (for example, cognitive demand, and habitability). They may therefore have recorded poorer perceptions towards, but better performance with the service. In contrast, the female participants may have perceived the service as being relatively easy to use, but not enough to cause any subsequent performance benefits. That male performance was better in public, but not in private, may be a function of male participant's tendency to perceive the service as being particularly useful in public places, such as when in a city, which resulted in enhanced levels of performance. There have been no previous studies that have found any effects of gender on attitudes towards the usability of non-metaphor standard automated telephone services. These findings therefore demonstrate the importance of considering gender as an important factor in the design of non-metaphor automated phone services.

Participants with relatively poor working memory perceived the standard service more positively than those with good working memory. It may have been the case that users with high working memory found the service easy to use, but found the actual user experience disappointing, which was reflected in their poor perceptions. There

have been few previous studies of working memory within HCI because the dominant GUI style rarely requires users to remember information, rather, the display acts as a visual reminder. There have been no previous studies examining the effect of working memory on speech systems, and as such, the work reported in this thesis demonstrates the importance of designing speech systems based on a thorough understanding of the limitations of working memory.

Participants with low levels of fixed line telephone usage perceived the service more positively. Inexperienced users of automated telephone services prefer a rigid, prompted interaction style (Zoltan-Ford, 1991), which is the style of interaction provided by the standard service. Such users may lack confidence in using the telephone, and prefer the non-conversational style of interaction presented by the standard service, which relies on prompts rather than conversational cues to maintain the interaction. These users may have found the simplicity of the number-based menus, which were the same at each level of the service, to be initially appealing. In contrast, this result suggests that experienced users were not provided with the kinds of features that enabled them to capitalise on their past experience and knowledge of the system, and became frustrated. There have been no previously documented studies that have analysed the effect of domain experience from fixed line telephones on users performance and attitudes towards mobile phone services. The finding outlined above therefore highlights the important role that domain experience can have on a user's ability to effectively use a new speech-based system.

In experiment three the extent to which services were visualised was assessed with reference to a definition of visualisation ability proposed by Ekstom et al. (1976). Visualisation was therefore taken as a measure of visualisation ability. Metaphor-based services were visualised significantly more often, and visualisation was found to have a significant effect on performance across all three of the services evaluated, both metaphor and non-metaphor, which may be attributed to the mental model formed as a result of the visualisation. This result supports and extends the findings of Vicente et al. (1987) who found visualisation ability to be a strong predictor of performance with a hierarchical file structure. Campagnoni and Ehrlich (1989) also found that users with high visualisation ability were better able to construct a mental

model of hierarchically presented information, which supports the construction of services that are designed to aid visualisation.

The individual difference results suggest that, if automated telephone services continue to be designed using numbered menu options, designers need to consider the ways in which gender, working memory, and telephone experience affect usability, and of ways that such factors may be accommodated through different designs. Alternatively, if designers choose to use interface metaphor, it may be possible to design services that will accommodate the effects of gender, working memory, and telephone experience, resulting in the design of metaphor-based phone services that are easier to use for a wider range of people.

8.1.3 Environment component: Context of use

Mobile phones may be used anywhere, which means that mobile phone services may be accessed from anywhere. In fact, the majority of mobile phone calls are made in busy public settings (Wei and Leung 1999). A study of the circumstances in which an interactive system will be expected to operate is a critical part of a HCD process (Preece et al., 2002). It is for this reason that an investigation of the effects of both physical and social context of use on service usability was conducted as part of experiment two, which was reported in chapter six of this thesis. Contexts of use were categorised as being private and public, with the actual locations within each being chosen by participants. This division of contexts was formulated as a means of comparing the physical and social aspects of mobile phone use in locations such as busy city streets, with that of less cognitively demanding locations, such as the home.

A quantitative analysis of attitude and performance data revealed that the context in which both the office filing system service and the standard service were used did not significantly impact usability. Subjective reports suggested that the visualisation strategy adopted by the majority of participants was still viable in conjunction with the visual and auditory distractions, and everyday physical navigation, which are features of operating such services in public contexts of use. Moreover, although participants usually reported at least one source of distraction when using services in public, most did not believe that their performance had been negatively impacted by it, which supports the quantitative results.

The social context of use caused social inhibitions for some participants, specifically those using the office filing system service, who often reported feeling slightly embarrassed or self-conscious when saying the service menu options. When using a metaphor-based service in public the social context may then influence how comfortable a user feels, a feature of mobile phone usage first studied Palen et al. (2001). One of their results, concerning the attitudes of the surrounding public towards the overheard phone conversation, may help to explain the effect of social context with respect to metaphor-based mobile phone services. They found that people who overheard phone conversations often perceived them to be frivolous or inconsequential. This was largely the result of them hearing only half of the conversation, and having no understanding of the context within which the conversation was occurring. During the experiments conducted for this thesis, people who overheard a participant's interaction with the metaphor-based service may then have perceived the 'conversation' as being not only frivolous or inconsequential, but also nonsensical. This is due to the abstract nature of the menu options spoken by participants, which would not have made sense to anyone overhearing the interaction, and which, if the user was aware of this, may subsequently have caused them to feel uncomfortable or embarrassed.

Due to different locations being represented by a single context of use category, either private or public, it is possible that participants may be able to use services more effectively in certain locations. However, it can be concluded that the broad contexts of use, defined for the work reported in this thesis, did not significantly affect usability, which provides evidence that speech-based services may provide a convenient alternative to screen-based information access for mobile users.

8.2 Limitations and future directions

Due to the limitations of time and resources that are an inherent aspect of any PhD, there is research work that could not be performed during this PhD, but which provides the basis for future research. The limitations of the work reported in this thesis, and the ways in which these limitations could be addressed by future work, will be discussed in this section with reference to three of the important components of HCI proposed with the Eason (1991) model: computer, human, and environment.

In terms of the 'computer' component, the extent to which single spatial metaphors, such as the office filing system, might be applicable to other more complex services needs to be investigated. A possible restriction on the degree to which a spatial metaphor could be applied to a more complex service is if that service allowed the user greater functionality than simply information retrieval, and an enhanced level of interactivity and flexibility. Services with large databases, such as the yellow pages, may also present problems. In such services, it is possible that a single metaphor may not be appropriate due to the potential difficulties of scaling spatial metaphors to cover the full system functionality and architecture of more complex systems. Under such circumstances, the use of multiple metaphors may need to be explored.

In preliminary study one, the results suggested that if a metaphor could be successfully applied to one automated mobile phone service, it could subsequently be successfully transferred to a different service domain. This finding was based on an analysis of metaphor preferences for two different mobile phone services, which revealed that seven of the ten metaphors rated as being applicable to a telephone Internet service were also rated as being applicable to a telephone city guide service. A useful starting point in this investigation would be to apply the office filing system metaphor to a number of different mobile phone service domains to assess its affect on their usability.

A future area of research relating to the 'computer' component is the evaluation of the office filing system metaphor service for location-based services. For the mobile city guide service to be effective, efficient, and satisfying to use, it would be necessary for the service to be able to detect the user's location and then to provide location specific information, possibly within user-defined parameters, such as, within a one hundred metre radius, or within a one-kilometre radius of their current location. It would also be necessary for the city guide service to be fully functional, rather than aspects of functionality being simulated, and for it to provide city information 24 hours a day. In so doing, the speech recognition aspect of the service could then also be evaluated, rather than being simulated using WOZ. For such functionality to be implemented, the metaphor may need to be extended, and a further investigation would serve to test the appropriateness of the metaphor to support additional functions.

For the 'human' component there are three main areas for future work. The first derives from the limitation of using participants who may not have been fully motivated to perform tasks with the service. If the experiments had been conducted with participants who were in a city, and actually needed the information they were attempting to access, perceptions towards the service may have been different. Future work could investigate this point by measuring perceptions towards a fully functional version of the city guide service, accessed through necessity, rather than obligation, and then gathering data about how participants act on the information they have accessed. An alternative approach would be to provide participants with less directed, scenario-based tasks. The scenarios would be based around, and extend, the locations from which participants accessed the service, thus helping participants to understand why they might want to use the service to perform the required task, increasing their perceived stake in the task, and leading to a more realistic, motivated interaction.

Good visualisation ability led to significantly improved performance with the services, but was only assessed in a self-reported way. Participants' descriptions of their service visualisations were compared against the definition of visualisation ability proposed by Ekstrom et al. (1976) in order to decide whether they had fulfilled the visualisation criteria. Future work could explore a more objective approach to the measurement of visualisation ability, the results from which could then form a quantitative measure of visualisation to be compared against the current qualitative metric. Gender, working memory, and telephone experience emerged as being significant predictors of attitude with the standard service, and gender as being a predictor of performance in public locations. It would be necessary to further investigate the reasons why these individual differences were important, and a qualitative study would enable a deeper level of understanding to be gained.

Thirdly, metaphors, and metaphor type, are culture specific, and a metaphor may have a high degree of salience to one culture, but be meaningless or irrelevant to another. A cross-cultural comparison of the usability of the office filing system metaphor would enable any inappropriate cultural aspects of the metaphor to be evaluated.

Three areas of future work relate to the 'environment' component. Firstly, due to the range of possible locations within both the private and public contexts that were

evaluated, it is possible that participants may be better able to visualise the service in certain locations. Investigating sub-locations within each of these broad categories may therefore be worthy of further study, for example, main road vs. street vs. pedestrian zone. Controlling the exact locations, in a similar way to that undertaken in experiment three, may offer an alternative approach, enabling factors such as background noise to be controlled. Secondly, there may be an experimental difference between receiving a call whilst performing another activity within a specific context of use, and taking part in an experiment in which the call is the main activity. Further investigation may therefore benefit from making calling times more random, to evaluate the usability of the services at times when a user may be multitasking, which may more accurately reflect real world usage. Finally, some participants appeared to have developed strategies for dealing with distractions and social inhibitions when using these services in public locations. Such strategies may include avoiding sources of background noise whilst using the services, or accessing the services in locations where other people cannot easily overhear the call. An observation-based study, or an interview-based study may allow such strategies to be investigated, and would provide designers with additional requirements for design.

8.3 Final reflections

This research was driven by the need to address the usability problems of speech-based automated mobile phone services. The approach taken was to use spatial interface metaphors, which, through their ability to promote an internal visualisation of a non-visual interface, significantly improved performance with, and tended to improve perceptions towards, the service evaluated. The cognitive mechanism of visualisation may also offer specific benefits to users of such services when in busy public places, through its ability to reduce the mental demands of the task. The approach taken appears to have been successful, with the use of spatial interface metaphor providing a real solution to a practical problem, with the potential to provide usable speech-based information access to mobile phone users wherever they are.

:: REFERENCES

References

ABOWD, G., and MYNATT, E., 2000, Charting past, present and future research in ubiquitous Computing. *ACM Transactions on Computer-Human Interaction*, **7**(1), 29-58.

ALLEN, R.B., 1987, Social processes and computing. *Unpublished paper* (Bell Communications Research).

ALLEVA, F., HUANG, X.D., HWANG, M.Y., and JIANG, L., 1998, Can continuous speech recognizers handle isolated speech? *Speech Communication*, **26**, 183-189.

ALLPORT, D., ANTONIS, B., and REYNOLDS, P., 1972, On the division of attention: a disproof of the single channel hypothesis. *Quarterly Journal of Experimental Psychology*, **24**, 225-235.

ALTY, J.A., KNOTT, R.P., ANDERSON, B., and SMYTH, M., 2000, A framework for engineering metaphor at the user interface. *Interacting with Computers*, **13**, 301-322.

ALTY, J., 1993, Co-operative working and multimedia in telecommunications, In *Proceedings of the RACE International Conference on Intelligence in Broadband Services in Networks*, (Brussels: Belgium), pp. 3-9.

ALTY, J.L., and KNOTT, R.P., 1997, Metaphor and interface design: extending the mapping model, In *Proceedings of the XVI European Annual Conference on Human Decision Making and Manual Control*, (Kassel: Germany), pp. 36-42.

ANDERSON, B., SMYTH, M., KNOTT, R.P., BERGAN, M., BERGAN, J., and ALTY, J.L., 1994, Minimising conceptual baggage: making choices about metaphor, In Proceedings of the Ninth Conference of the British Computer Society HCI Specialist Group, Cambridge: CUP, pp. 179-194.

APPLE COMPUTER CORPORATION, 1985, *Apple Human Interface Guidelines. The Apple Desktop Interface* Reading, MA: Addison-Wesley.

APPLE COMPUTER CORPORATION, 1987, *Apple Human Interface Guidelines.* Reading, MA: Addison-Wesley.

APPLE COMPUTER CORPORATION, 1992, *Macintosh Human Interface Guidelines.* Reading, MA: Addison-Wesley.

ARISTOTLE, 1995, *Treatise on Rhetoric* (T. Buckley, Trans.). (New York: Prometheus Books). (Original work written ca. 328BC).

ARONSON, E., CARLSMITH, J.M., 1986, Experimentation in social psychology. In: *Handbook of Social Psychology*, edited by G. Lindzey, and E. Aronson (Reading, Mass: Addison-Wesley).

ATKINSON, R.L., ATKINSON, R.C., and HILGARD, E.R., 1983, *Introduction to Psychology* (Harcourt Brace Jovanovich).

AUCELLA, A.F., and EHRLICH, S.F., 1986, Voice messaging enhancing the user interface design based on field performance, In Proceedings of the SIGCHI Conference on Human Factors in computing Systems (Boston, Massachusetts: United States), pp.156-161.

AUCELLA, A.F., KINKEAD, R., SCHMANDT, C. and WICHANSKY, A., 1987, Voice: technology searching for communication needs, In Proceedings of the Conference on Human Factors in Computing Systems and Graphic Interfaces, pp. 41-44.

- BABER, C., 1993, Developing interactive speech technology. In: *Interactive Speech Technology*, edited by C. Baber, and J. Noyes, (London: Taylor and Francis).
- BADDELEY, A., and HITCH, G., 1974, Working memory. In: *Recent advances in learning and motivation*, edited by G.A. Bower (New York: Academic Press).
- BADDELEY, A., 1986, *Working memory* (Oxford: Oxford University Press).
- BADDELEY, A., 1992, Is working memory working? *Quarterly Journal of Experimental Psychology*, **44**(1), 1-31.
- BARNARD, E., HALBERSTADT, A., KOTELLY, C., and PHILLIPS, M., 1999, A consistent approach to designing spoken-dialog systems, In Proceedings of the Automatic Speech Recognition and Understanding Workshop (Keystone, CO: United States).
- BARREAU, D., and NARDI, B., 1995, Finding and reminding: file organization from the desktop. *ACM SIGCHI Bulletin*, **27**(3), 39-43.
- BASILI, V., CALDIERA, G., and ROMBACH, D.H., 1994, *The Goal Question Metric Paradigm: Encyclopedia of Software Engineering* (New York: John Wiley & Sons).
- BAUM, F., 1900, *The Wizard of Oz* (Collins, London).
- BEACHAM, K., and BARRINGTON, S., 1996, CallMinder – the development of BTs new telephone answering service. *BT Technology Journal*, **4**(2), 52-59.
- BECK, E., CHRISTIANSEN, M., KJELDSKOV, J., KOLBE, N., and STAGE, J., 2003, Experimental evaluation of techniques for usability testing of mobile systems in a laboratory setting, In Proceedings of OzCHI 2003 (Brisbane: Australia).

BENBASAT, I., and TODD, P., 1993, An experimental investigation of interface design alternatives: icon vs. text and direct manipulation vs. menus. *International Journal of Man-Machine Studies*, **38**(3), 369–402.

BENKING, H., and JUDGE, A., 1994, Design considerations for spatial metaphors, Presentation at The European Conference on Hypermedia Technology (Edinburgh: UK).

BENYON, D.R., and HOOK, K., 1997, Navigation in information spaces: supporting the individual, In *Proceedings of Human-Computer Interaction, INTERACT 1997*, edited by S. Howard, J. Hammond and G. Linedgaard (Chapman and Hall), pp. 39-46.

BENYON, D.R., 1993, Accommodating individual differences through an adaptive user interface. In: *Adaptive User Interfaces - Results and Prospects*, edited by M. Schneider-Hufschmidt, T. Kühme, and U. Malinowski (Elsevier Science Publications, North-Holland: Amsterdam).

BEYER, H., and HOLTZBLATT, K., 1998, *Contextual design: defining customer-centred systems* (San Francisco, California: Morgan Kaufmann).

BILLI, R., CANAVESIO, R., and RULLENT, C., 1998, Automation of telecom italia directory assistance service: field trial results, In *Proceedings of Interactive Voice Technology for Telecommunication Applications Conference*.

BILLINGHURST, M., and WEGHORST, S., 1995, The use of sketch maps to measure cognitive maps of virtual environments, In *Proceedings of IEEE Virtual Reality Annual International Symposium* (Los Alamitos: CA), pp. 40-47.

BLACK, M., 1962, *Models and Metaphors* (New York: Cornell University Press).

BLACKWELL, A.F., 2001, Pictorial representation and metaphor in visual language design. *Journal of Visual Languages and Computing*, **12**(3), 223-252.

- BLATTNER, M.M., SUMIKAWA, D.A., and GREENBERG, R.M., 1989, Earcons and icons: their structure and common design principles. *Human Computer Interaction*, **4**(1), 11-44.
- BLUMENTHAL, B., 1990, Strategies for automatically incorporating metaphoric attributes in interface designs, In Proceedings of the 3rd Annual ACM SIGGRAPH Symposium on User Interface Software and Technology (Snowbird, Utah: United States), pp. 66–75.
- BOND, C., and CAMACK, M., 1999, Your call is important to us ... please hold. *Ergonomics in Design*, **7**(4), 9-15.
- BORGMAN, C.L., 1986, The user's mental model of an information retrieval system: an experiment on a prototype online catalogue. *International Journal of Man-Machine Studies*, **24**, 47-64.
- BORGMAN, C.L., 1989, All users of information retrieval systems are not created equal: an exploration into individual differences. *Information Processing and Management*, **25**, 237-251.
- BOYD, R., 1979, Metaphor and theory change: what is “metaphor” a metaphor for? In: *Metaphor and Thought*, edited by A. Ortony (Cambridge: Cambridge University Press), pp. 481-532.
- BRAJNIK, G., GUIDA, G., and TASSO, C., 1990, User modeling in expert man-machine interface: a case study in intelligent information retrieval. *IEEE Transactions on Systems, Man and Cybernetics*, **20**, 166-185.
- BREAKWELL, G., HAMMOND, S.M., and FIFE-SCHAW, C., 2000, *Research methods in Psychology*, 2nd edition, (London, Sage).
- BREMS, D.J., RABIN, M.D., and WAGGETT, J.L., 1995, Using natural language conventions in the user interface design of automatic speech recognition systems. *Human Factors*, **37**(2), 265-282.

BREWSTER, S.A., RATY, V.P., and KORTEKANGAS, A., 1996, Earcons as a method of providing navigational cues in a menu hierarchy, In Proceedings of BCS HCI 1996 (London: UK), pp. 167-183.

BREWSTER, S.A., 1997, Navigating telephone-based interfaces with earcons, In Proceedings of BCS HCI 1997 (Bristol: UK), pp. 39-56.

BREWSTER, S.A., 2002, Overcoming the lack of screen space on mobile computers. *Personal and Ubiquitous Computing*, **6**, 188 – 205.

BREWSTER, S.A., WRIGHT, P.C., and EDWARDS, A.D.N., 1993, An evaluation of earcons for use in auditory human-computer interfaces, In Proceedings of InterCHI 1993 (Amsterdam: The Netherlands), pp. 222-227.

BRISTOW, G., 1986, The speech recognition problem. In: *Electronic Speech Recognition*, edited by G. Bristow (Collins).

BRYCE, A., 2000, Individual differences and the conundrums of user-centred design: two experiments. *Journal of the American Society for Information Science*, **6**, 508-520.

CAMPAGNONI, F.R., and EHRLICH, K., 1989, Information retrieval using hypertext-based help system. *ACM Transactions on Information Systems*, **7**(3), 271-291.

CAPLAN, L.J., and SCHOOLER, C., 1990, The effects of analogical training models and age on problem-solving in a new domain, *Experimental Aging Research*, **16**(3), 151-154.

CARD, S.K., MACKINLAY, J.D., and SHNEIDERMAN, B., 1999, *Readings in information visualization: using vision to think* (San Francisco, California: Morgan-Kaufmann).

CARROLL, J.B., 1993, *Human cognitive abilities: a survey of factor-analytic studies* (New York: Cambridge University Press).

CARROLL, J.M., MACK, R.L., and KELLOGG, W.A., 1988, Interface metaphors and user interface design. In: *Handbook of Human-Computer Interaction*, edited by M. Helander, (Elsevier Science Publishers: North Holland), pp. 67-85.

CARROLL, J.M., and MACK, R.L., 1985, Metaphor, computing systems and active learning. *International Journal of Man-Machine Studies*, **22**(1), 39-57.

CARROLL, J.M., and OLSON, J.R., 1988, Mental models in human-computer interaction. In: *Handbook of Human-Computer Interaction*, edited by M. Helander, (Elsevier Science Publishers: North Holland), pp. 67-85.

CARROLL, J.M., and THOMAS, J.C., 1982, Metaphor and the cognitive representation of computing systems. *IEEE Transactions on Systems, Man and Cybernetics*, **12**(2), 107-116.

CASALI, S.P., WILLIGES, B.H., and DRYDEN, R.D., 1990, Effects of recognition accuracy and vocabulary size of a speech recognition system on task performance and user acceptance. *Human Factors*, **32**(2), 183-196.

CATELL, R.B., EBER, H.W., and TATSUOKA, M.M., 1970, *Handbook for the sixteen personality factor questionnaire* (Champaign IL: Institute for Personality and Ability Testing).

CATES, W.M., 2002, Systematic selection and implementation of graphical user interface metaphors, *Computers and Education*, **38**, 2002, 385-397.

CHAMBERS, R.M., and DE HAAN, H.J., 1985, Applications of automated speech technology to land-based army systems. *Speech Technology*, (February/March), 92-99.

CHAN, K.F., and YEUNG, D.Y., 1999, Recognizing on-line handwritten alphanumeric characters through flexible structural matching. *Pattern Recognition*, **32**, 1099 -1114.

CHAVAN, A.L., 1994, A design solution project on alternative interface for MS Windows. Master's thesis, London Guildhall University (London: United Kingdom).

CHEN, C., CZERWINSKI, M., and MACREDIE, R., 2000, Individual differences in virtual environments – Introduction and overview. *Journal of the American Society for Information Science*, **51**(6), pp. 499-507.

CHEN, C., and RADA, R., 1996, Interacting with hypertext: a meta-analysis of experimental studies. *Human-Computer Interaction*, **11**(2), 125-156.

CHOINERE, A., ROBERT, J.M., and DESCOUT, R., 1991, Building a user interface for a speech-based telephone application system, In Proceedings of Eurospeech 1991, pp. 1503-1506.

CHRYSLER, E., 1978, Some basic determinants of computer programming productivity. *Communications of the ACM (CACM)*, **21**(6), 472-483.

COLLINS, D., 1995, *Designing object-oriented user interfaces* (Redwood City, CA: Benjamin/Cummings Publishing Company, Inc).

COLOM, R., CONTRERAS, M.J., BOTELLA, J., and SANTACREU, J., 2001, Vehicles of spatial ability. *Personality and Individual Differences*, **32**(5), 903-912.

CONDON, C., and KEUNEKE, S., 1994, Metaphors and layers of signification: the consequences for advanced user service interfaces, In Proceedings of RACE IS & N Conference, 7-9 September (Aachen: Germany), pp. 75-87.

COOLICAN, H., 1990, *Research Methods and Statistics in Psychology* (Hodder and Stoughton).

CORNELL WAY, E., 1991, *Metaphor and knowledge representation* (Kluwer).

COVENTRY, L., 1989, Some effects of cognitive style on learning UNIX. *International Journal of Man-Machine Studies*, **31**(3), 349-365.

CROWDER, R.G., 1970, The role of one's own voice in immediate memory. *Cognitive Psychology*, **1**, 157-178.

CROWDER, R.G., 1976, *Principles of learning and memory* (Hillsdale, NJ: Erlbaum).

DAMPER, R.I., 1993, Speech as an interface medium: how can it best be used? In: *Interactive Speech Technology*, edited by C. Baber, J.M. Noyes (London: Taylor and Francis).

DANCEY, C.P., and REIDY, J., 2002, *Statistics Without Maths for Psychology: Using SPSS for Windows™*, 2nd edition (London: Prentice Hall).

DAVIS, F.D., BAGOZZI, R.P., and WARSHAW, P.R., 1989, User acceptance of computer technology: a comparison of two theoretical models. *Management Science*, **35**, 982-1002.

DE BONO, E., 1992, *Serious Creativity* (Harper Business: New York, US).

DEVAUCHELLE, P., 1991, *User-friendly recommendations for voice services designers* (France Telecom: NT/LAA/TSS/426 – TARIF: 150 F HT (177, 90 TTC)).

DEY, A., ABOWD, G., and SALBER, D., 2001, A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human-Computer Interaction*, **16**(2-4), 97-166.

DIEBERGER, A., 1994a, Spatial environments to organize and navigate information and to communicate about this organization, Position Paper at the ECHT'94 Workshop on Spatial Metaphors in Hypermedia Systems.

DIEBERGER, A., 1994b, Navigation in textual virtual environments using a city metaphor, PhD Thesis, (Vienna University of Technology).

DIEBERGER, A., 1996, Browsing the WWW by interacting with a textual virtual environment - a framework for experimenting with navigational metaphors, In Proceedings of Hypertext 1996 (Washington DC: United States), pp. 170-179.

DIEBERGER, A., and FRANK, U., 1998, A city metaphor to support navigation in complex information spaces. *Journal of Visual Languages and Computing*, **9**, 597-622.

DILLON, A., RICHARDSON, J. and MCKNIGHT, C., 1999, The effect of display size and text splitting on reading lengthy text from the screen. *Behaviour and Information Technology*, **9**(3), 215-227.

DILLON, A., and WATSON, C., 1996, User analysis in HCI - the historical lessons from individual differences research. *International Journal of Human-Computer Studies*, **45**(6), 619-637.

DINTRUFF, D.L., GRICE, D.G., and WANG, T.G., 1985, User acceptance of speech technologies. *Speech Technology*, **2**(4), 16-21.

DIX, A., FINLAY, J., ABOWD, G., and BEALE, R., 2004, *Human-Computer Interaction*, 3rd edition (Prentice Hall).

DIX, A., RODDEN, T., DAVIES, N., TREVOR, J., FRIDAY, A., and PALFREYMAN, K., 2000, Exploiting space and location as a design framework for interactive mobile systems. *ACM Transactions in Computer-Human Interaction*, **7**(3), 285-321.

DOUGLAS, S.A., and MORAN, T.P., 1983, Learning text editor semantics by analogy, In Proceedings of CHI 1983: Human Factors in Computing Systems (New York: ACM Publications).

DOURISH, P., 2004, What we talk about when we talk about context. *Personal and Ubiquitous Computing*, **8**(1), 19-30.

DREISTADT, R., 1968, An analysis of the use of analogies and metaphors in science. *The Journal of Psychology*, **68**, 97-116.

DU BOULAY, B., O'SHEA, T., and MONK, J., 1981, The black box inside the glass box: presenting computer concepts to novices. *International Journal of Man-Machine Studies*, **14**, 237-249.

DUMAIS, S.T., and WRIGHT, A.L., 1986, Reference by name vs. location in a computer filing system, In Proceedings of the Human Factors Society, pp. 824-828.

DUMAS, J.S., and REDISH, J.C., 1999, *A practical guide to usability testing*, 2nd edition (Intellect).

DUNCAN, S., 1972, Some signals and rules for taking speaking turns in conversation, *Journal of Personality and Social Psychology*, **23**, 282-292.

DURFRESNE, A., and TURCOTTE, S., 1997, Cognitive style and its implications for navigation strategies. In: *Artificial intelligence in education knowledge and media learning system*, edited by B. Boulay and R. Mizoguchi (Amsterdam IOS Press), pp. 287-293.

DUTTON, R., FOSTER, J. and JACK, M., 1999, Please mind the doors – do interface metaphors improve the usability of voice response services? *BT Technology Journal*, **17**(1), 172-177.

EASON, K.D., 1984, Towards the experimental study of usability. *Behaviour and Information Technology*, **3**, 133-143.

EASON, K.D., 1991, Ergonomic perspective on advances in human-computer interaction, *Ergonomics*, **34**, 721-741.

- EDMAN, T.R., and METZ, S.V., 1983, A methodology for the evaluation of real-time speech digitization, In Proceedings of the Human Factors Society 27th Meeting, pp. 104-107.
- EDWARDS, K.W., and STOCKTON, K., 1995, Access to GUIs for blind users, *Interactions*, 2(1), 54-67.
- EGAN, D.E., and GOMEZ, L.M., 1985, Assaying, isolating, and accommodating individual differences in learning a complex skill. In: *Individual Differences in Cognition*, edited by R.F. Dillon, (Orlando: Academic Press), pp. 173-217.
- EGAN, D., 1988, Individual differences in human-computer interaction. In: *Handbook of Human-Computer Interaction*, edited by M. Helander, (Elsevier Science Publishers: North Holland), pp. 543-568.
- EKSTROM, R.B., FRENCH, J.W., HARMAN, H.H., and DERMEN, D., 1976, *Manual for kit of factor-referenced cognitive tests* (Educational Testing Service, Princeton: NJ).
- ELIAS, P.K., ELIAS, M.F., ROBBINS, M.A., and GAGE, P., 1987, Acquisition of word processing skills by younger, middle age, and older adults. *Psychology and Ageing*, 2, 340-348
- ELKERTON, J., and WILLIGES, R.C., 1984, Information retrieval strategies in a file search environment, *Human Factors*, 26, 171-184.
- ENGELBECK, G., and ROBERTS, T., 1989, The effects of several voice-menu characteristics on menu-selection performance. *U.S. West Advanced Technologies Technical Report*, 119.
- ENGLE, W., 1974, Modality effect: is pre-categorical acoustic storage responsible? *Journal of Experimental Psychology*, 102, 824-829.

- FISCHER, G., 1991, The importance of models in making complex systems comprehensible. In: *Mental Models and Human-Computer Interaction 2*, edited by M.J. Tauber and D. Ackermann (Amsterdam: Elsevier Science Publishers, BV, North-Holland).
- FORD, N., and CHEN, S.Y., 2000, Individual differences, hypermedia navigation and learning: an empirical study, *Journal of Educational Multimedia and Hypermedia*, 9(4), 281–312.
- FOSS, D.J., ROSSON, M.B., and SMITH, P.L., 1982, Reducing manual labor: an experimental analysis of learning aids for a text editor, In Proceedings of Human Factors in Computing Systems Conference, CHI 1982 (New York: ACM), pp. 423-429.
- FOSTER, J.C., MCINNES, F.R., JACK, M.A., LOVE, S., DUTTON, R.T., NAIRN, I.A., and WHITE, L.S., 1998, An experimental evaluation of preferences for data entry method in automated telephone services, *Behaviour and Information Technology*, 17(2), 82-92.
- FOWLER, C.J.H., MACAULAY, L.A, and SIRIPOKSUP, S., 1988, An evaluation of the effectiveness of the adaptive interface module (AIM) in matching dialogues to users. In: *People & Computers III*, edited by D. Diaper and R. Winder (Cambridge: CUP).
- FRASER, N.M., and GILBERT, G.N., 1991, Simulating speech systems. *Computer Speech and Language*, 5(1), 81-99.
- FREKSA, C., 1991, Qualitative spatial reasoning. In: Cognitive and linguistic aspects of geographical space, edited by D. Mark, and A. Frank A. (Kluwer), pp. 361-372.
- FU, H.C., CHANG, H.Y., XU, Y.Y., and PAO, H.T., 2000, User adaptive handwriting recognition by self-growing probabilistic decision-based neural networks. *IEEE Transactions on Neural Networks*, 11, 1373-1384.

GALLWITZ, F., NIEMANN, H.& NOTH, E. (1999). Speech recognition - state of the art, applications, and future prospects. *Wirtschaftsinformatik*, **41**, 538-549.

GARDEN, S., CORNOLDI, C., and LOGIE, R.H., 2002, Visuo-spatial working memory in navigation. *Applied Cognitive Psychology*, **16**, 35–50.

GATTIKER, U., and HLAVKA, A., 1992, Computer attitudes and learning performance: issues for management education and training. *Journal of Organizational Behavior*, **13**, 89-101.

GAVER, W.W., 1989, The SonicFinder: an interface that uses auditory icons. *Human Computer Interaction*, **4**(1), 67-94.

GAVER, W.W., 1995, Oh what a tangled web we weave: metaphor and mapping in graphical interfaces, In Proceedings of ACM CHI 1995 (Denver, CO: USA).

GENTNER, D., and NIELSEN, J., 1996, The anti-Mac interface. *Communications of the ACM*, **39**(8), 70-82.

GENTNER, D., 1983, Structure-mapping: a theoretical framework for analogy. *Cognitive Science*, **7**, 155-170.

GENTNER, D., and GENTNER, D.R., 1983, Flowing waters or teeming crowds: mental models of electricity. In: *Mental Models*, edited by D. Gentner, and A. Stevens (Lawrence Erlbaum Press).

GENTNER, D., and JEZIORSKI, M., 1993, From metaphor to analogy in western science. In: *Metaphor and Thought*, 2nd edition, edited by A. Ortony (Cambridge: Cambridge University Press), pp. 447-480.

GILLAN, D.J., FOGAS, B.S., ABERASTURI, S., and RICHARDS, S., 1995, Cognitive ability and computing experience influence interpretation of computer metaphors, In Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting, pp. 243-247.

GOFFMAN, E., 1959, *The Presentation of Self in Everyday Life*. (Doubleday: Garden City, New York).

GOMEZ, L.M., EGAN, D.E., and BOWERS, C., 1986, Learning to use a text editor: some learner characteristics that predict success. *Human-Computer Interaction*, **2**, 1-23.

GOODING, D., 1996, Diagrams in the generation and dissemination of new science: some examples and applications. In Proceedings of the IEE Colloquium on Thinking with Diagrams (London: IEE Computing and Control Division), pp. 3/1-3/6.

GOULD, J., and BOIES, S.J., 1983, Human factors challenges in creating a principal support office system - the speech filing approach. *ACM Transactions on Office Information Systems*, **1**(4), 273-298.

GOULD, J.D., BOIES, S.J., LEVY, S., RICHARDS, J.T., and SCHOONARD, J., 1987, The 1984 Olympic message system: a test of behavioral principles of system design. *Communications of the ACM*, **30**(9), 758-769.

GOULD, J., CONTI, J., and HOVANYECZ, T., 1983, Composing letters with a simulated listening typewriter. *Communications of the ACM*, **26**(4), 295-308.

GREEN, T.R.G., and BENYON, D.R., 1996, The skull beneath the skin: entity-relationship modelling of information artefacts. *International Journal of Human-Computer Studies*, **44**(6), 801-828.

GREEN, T.R.G., and NAVARRO, R., 1995, Programming plans, imagery, and visual programming, In Proceedings of INTERACT 1995 (London: Chapman and Hall), pp. 139-144.

GREENE, S., GOMEZ, L., and DEVLIN, S., 1986, A cognitive analysis of database query production, In Proceedings of the Human Factors Society 30th Annual Conference (Santa Monica, CA: United States), pp. 9-13.

GUYOMARD, M., and SIROUX, J., 1988, Experimentation in the specification of an oral dialogue. In: *Recent Advances in Speech Understanding and Dialog Systems*, edited by H. Niemann, M. Lang, and G. Sagerer (NATO ASI Series F), **46**, pp. 497-502.

HALASZ, F.G., and MORAN, T.P., 1982, Analogy considered harmful, In *Proceedings of the Conference on Human Factors in Computing Systems* (Gaithersburg, MD: United States), pp. 383-386.

HALSTEAD-NUSSLOCH, R., 1989, The design of phone-based interfaces for consumers, In *Proceedings of CHI 1989 Human Factors in Computing Systems* (Austin, Texas: United States), pp. 347-352.

HAMILTON, A., 2000a, Interface metaphors and logical analogues: a question of terminology. *Journal of the American Society for Information Science*, **51**(2), pp.111-122.

HAMILTON, A., 2000b, Metaphor in theory and practice: the influence of metaphors on expectations, *ACM Journal of Computer Documentation*, **24**(4), pp. 237–253.

HAMMOND, N., and ALLINSON, L., 1987, The travel metaphor as design principle and training aid for navigating around complex systems, In *Proceedings of HCI 1987*, pp.75-90.

HAN, S.H., and KWAHK, J., 1994, Design of a menu for small displays presenting a single item at a time, In *Proceedings of the Human factors and Ergonomics Society 38th Annual Meeting*, (Nashville: USA), pp. 360–364.

HAPESHI, K., and JONES, D.M., 1988, The ergonomics of automatic speech recognition interfaces. In: *International Reviews of Ergonomics*, edited by D.J. Osborne, pp. 251-290.

- HAPESHI, K., 1993, Design guidelines for using speech in interactive multimedia systems. In: *Interactive Speech Technology*, edited by C. Baber, and J. Noyes (London: Taylor & Francis), pp. 177-188.
- HARRISON, B.L., FISHKIN, K.P., GUJAR, A., MOCHON, C., and WANT, R., 1998, Squeeze me, hold me, tilt me! An exploration of manipulative user interfaces, In Proceedings of ACM CHI 1998 (Los Angeles, CA: United States), pp. 17-24.
- HAUPTMANN, A.G., 1989, Speech and gestures for graphic image manipulation, In Proceedings of CHI 1989 (Austin, Texas: United States), pp. 241-245.
- HEIM, A.W., 1970, *AH4 Group Test of General Intelligence* (UK: NFER).
- HOBBS, T., 1914, *The Leviathan* (London: Dent and Dutton). (Original work published London, 1651).
- HONE, K.S., and GRAHAM, R., 2000, Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Natural Language Engineering*, 6(3/4), 287-305.
- HUTCHINS, E., 1989, Metaphors for interface design. In: *The structure of multimodal dialogues*, edited by M.M. Taylor, F. Neel, and D.G. Bouwhuis (Amsterdam: North-Holland), pp. 11-28.
- INDURKHYA, B., 1992, *Metaphor and Cognition* (Boston: Kluwer Academic Publishers).
- ISO 13407:1999, *Human-centred design processes for interactive systems*.
- ISO 9241:1998-11, *Ergonomics of office work with VDTs - Guidance on usability*.
- JAFFE, J., and FELDSTEIN, S., 1970, *Rhythms of dialogue* (London: Academic Press).

- JENNINGS, F., BENYON, D., and MURRAY, D., 1991, Adapting systems to differences between individuals. *Acta Psychologica*, **78**(1-3), 243-258.
- JOHNSON, J., 1987, How faithfully should the electronic office simulate the real one? *SIGCHI Bulletin*, **19**(2), 21-25.
- JOHNSON, P., 1998, Usability and mobility: Interactions on the move, In Proceedings of the First Workshop on Human-Computer Interaction with Mobile Devices (Glasgow: Scotland), GIST Technical Report G98-1.
- JOHNSON-LAIRD, P., 1983, *Mental Models* (Cambridge, MA: Harvard University Press).
- JOHNSON-LAIRD, P.N., 1988, Freedom and constraint in creativity. In: *The nature of creativity: Contemporary psychological perspectives*, edited by R.J. Sternberg (Cambridge: Cambridge University Press), pp. 202-219.
- JONASSEN, D., and GRABOWSKI, B.L., 1993, *Handbook of Individual Differences. Learning and Instruction* (Lawrence Erlbaum Associates, Inc).
- JONASSEN, D., and WANG, S., 1993, Acquiring structural knowledge from semantically structured hypertext. *Journal of Computer-based Instruction*, **20**(1), 1-8.
- JONES, A., 1982, Mental Models of a first programming language. *CAL Research Group Technical Report No. 29*.
- JONES, D.M., FRANKISH, C.R., and HAPESHI, K., 1992, Automatic speech recognition in practice. *Behaviour and Information Technology*, **11**, 109-122.
- JONES, D.M., HAPESHI, K., and FRANKISH, C.R., 1989, Design guidelines for speech recognition interfaces, *Applied Ergonomics*, **20**, 47-52.

JONES, D.M., MILES, C., and PAGE, J., 1989, Disruption of reading by irrelevant speech: Effects of memory, arousal, or attention? *Journal of Applied Cognitive Psychology*, **4**, 89-108.

JONES, J. C., 1980, *Design Methods: Seeds of Human Futures* (Chichester: Wiley).

JONES, M., MARSDEN, G., MOHD-NASIR, N., BOONE, K. and BUCHANAN, G., 1999, Improving web interaction on small displays, In Proceedings of the WWW8 Conference (Toronto: Canada), pp. 51–59.

JUNGK, A., THULL, B., FEHRLE, L., HOEFT, A., and RAU, G., 2000, A case study in designing speech interaction with a patient Monitor. *Journal of Clinical Monitoring and Computing*, **16**(4), 295-307.

KAHN, R., and CANNELL, C., 1957, *The Dynamics of Interviewing* (Wiley: New York).

KAMM, C., and HELANDER, M., 1997, Design issues for interfaces using voice input. In: *Handbook of Human Computer Interaction*, edited by M. Helander, T.K. Landauer, and P. Prabhu (Elsevier Science Publishers), pp. 1043-1059.

KAPTELININ, V., 1996, Creating computer-based work environments: an empirical study of Macintosh users, In Proceedings of the 1996 ACM SIGCPR/SIGMIS Conference on Computer Personnel Research, pp. 360–366.

KARIS, D., and DOBROTH, K., 1991, Automating services with speech recognition over the public switched telephone network: human factors considerations. *IEEE Journal on Selected Areas in Communications*, **9**(4), pp. 574-585.

KAY, A., 1990, User interface: a personal view. In: *The Art of Human Computer Interface Design*, edited by B. Laurel, and S. Mountford (New York: Addison-Wesley Inc).

- KELLEY, C.L., and CHARNESS, N., 1995, Issues in training older adults to use computers. *Behaviour and Information Technology*, 14(2), 107-120.
- KIDD, A., 1985, Problems of man-machine dialogue design, In Proceedings of the 6th International Conference on Computer Communications, pp. 531-536.
- KIM, H., and HIRTLE, S.C., 1995, Spatial metaphors and disorientation in hypertext browsing. *Behaviour & Information Technology*, 14, 239-250.
- KIM, J., 1999, An empirical study of navigation aids in customer interfaces. *Behaviour and Information Technology*, 18(3), 213-224.
- KIRAKOWSKI, J., 1996, The software usability measurement inventory: background and usage. In: *Usability Evaluation in Industry*, edited by P.W. Jordan, B. Thomas, B.A. Weerdmeester, and I.L. McClelland (London: Taylor & Francis), pp. 169-178.
- KITTAY, E., 1987, *Metaphor: Its Cognitive Force and Linguistic Structure* (Oxford: Clarendon Press).
- KJELDSKOV, J., 2002, Just-in-place information for mobile device interfaces, In Proceedings of Mobile HCI 2002 (Pisa: Italy), Lecture Notes in Computer Science (Berlin, Springer-Verlag), pp. 271-275.
- KJELDSKOV, J., and SKOV, M.B., 2003, Creating a realistic laboratory setting: a comparative study of three think-aloud usability evaluations of a mobile system, In Proceedings of the 9th IFIP TC13 International Conference on Human Computer Interaction, Interact 2003 (Zürich: Switzerland).
- KJELDSKOV, J., SKOV, M.B., ALS, B.S., and HØEGH, R.T., 2004, Is it worth the hassle? Exploring the added value of evaluating the usability of context-aware mobile systems in the field, In Proceedings of the 6th International Mobile HCI 2004 conference (Glasgow: Scotland), Lecture Notes in Computer Science (Berlin, Springer-Verlag).

KLEMMER, S.R., SINHA, A.K., CHEN, J., LANDAY, J.A., ABOOBAKER, N., and WANGET, A., 2000, SUEDE: A Wizard of Oz prototyping tool for speech user interfaces, In Proceedings of ACM Symposium on User Interface Software and Technology, pp. 1-10.

KLERER, M., 1984, Experimental study of a two-dimensional language vs. FORTRAN for first course programmers. *International Journal of Man-Machine Studies*, **20**, 573-592.

KOESTLER, A., 1964a, *The Sleepwalkers: A History of Man's Changing Vision of the Universe* (London: Penguin).

KOESTLER, A., 1964b, *The act of creation* (London: Hutchinson).

KOUBEK, R.J., LEBOULD, W.K., and SALVENDY, G., 1985, Predicting performance in computer programming courses, *Behaviour and Information Technology*, **4**(2), 13-129.

KUHN, W., and FRANK, A.U., 1991, A formalization of metaphors and image-schemas in user interfaces. In: *Cognitive and linguistic aspects of geographic space*, edited by D.M. Mark, and A.U. Frank (Kluwer), pp. 419-434.

KUHN, W., and BLUMENTHAL, B., 1996, Spatialization: spatial metaphors for user interfaces, In Proceedings of CHI 1996 (Vancouver, BC: Canada).

LABRADOR, C., and DINESH, P., 1984, Experiments in speech interaction with conventional data services, In Proceedings of INTERACT 1984 (London: UK), pp. 104-108.

LAKOFF, G., 1993, The contemporary theory of metaphor. In: *Metaphor and Thought*, edited by A. Ortony, 2nd edition (Cambridge University Press), pp. 202-251.

LAKOFF, G., and JOHNSON, M., 1980, *Metaphors We Live By* (Chicago: University of Chicago Press).

LARKIN, J.H., 1983, Problem representations in physics. In: *Mental Models*, edited by A.L. Stevens, and D. Gentner (Hillsdale, NJ: Lawrence Erlbaum), pp. 75-98.

LARKIN, J.H., and SIMON, H.A., 1987, Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, **11**(1), 65-99.

LAUREL, B., 1990, Interface agents: Metaphors with character. In: *The Art of human computer interface design*, edited by B. Laurel (MA: Addison-Wesley).

LEA, W.A., 1980, The value of speech recognition systems. In: *Trends in Speech Recognition*, edited by W.A. Lea (Prentice Hall: Englewood Cliffs, NJ).

LEEDHAM, C.G., 1991, Speech and handwriting. In: *Engineering the Human-Computer Interface*, edited by A.C. Downton (McGraw Hill), pp. 220-245.

LEGGETT, J., and WILLIAMS, G., 1984, An empirical investigation of voice as an input modality for computer programming. *International Journal of Man - Machine Studies*, **21**, 493 - 520.

LEPLATRE, G., and BREWSTER, S.A., 2000, Designing Non-Speech Sounds to Support Navigation in Mobile Phone Menus, In Proceedings of ICAD 2000 (Atlanta: USA), pp. 190-199.

LIEBERMAN, H., 1997, Autonomous Interface Agents, In Proceedings of the Conference on Computers and Human Interface, CHI 1997, pp. 67-74.

LIKERT, R.A., 1932, A technique for the measurement of attitudes. *Archives of psychology*, **140**, pp. 55.

LING, R., 1996, The technological definition of social boundaries: video telephony and the constitution of group membership. *Teletronikk*, **1**, 61-73.

LIU, M., and REED, W.M., 1995, The effect of hypermedia assisted instruction on second-language learning through a semantic-network-based approach. *Journal of Educational Computing Research*, **12**(2), 159–175.

LOHMAN, D.F., 1989, Human intelligence: an introduction to advances in theory and research, *Review of Educational Research*, **59**(4), 333–373.

LOHMAN, D.F., PELLEGRINO, J.W., ALDERTON, D.L., and REGIAN, J.W., 1987, Dimensions and components of individual differences in spatial abilities. In: *Intelligence and cognition: Contemporary frames of reference*, edited by S.H. Irvine, and S.E. Newstead (Dordrecht, The Netherlands: Martinus Nijhoff), pp. 253-312.

LOVE, S., 1997, The role of individual differences in dialogue engineering for automated telephone services, *Unpublished PhD thesis* (University of Edinburgh: UK).

LOVE, S., and PERRY, M., 2004, Dealing with mobile conversations in public places: some implications for the design of socially intrusive technologies, In *Proceedings of CHI 2004* (Vienna: Austria), pp. 1195-1198.

LOVE, S., FOSTER, J., and JACK, M., 1997, Assaying and isolating individual differences in automated telephone services, In *Proceedings of the 16th International Symposium on Human factors in Telecommunications (HFT'97)*, pp. 323-330.

LOVE, S., FOSTER, J.C., and JACK, M.A., 2000, Health warning: use of speech synthesis can cause personality changes, *State of the Art in Speech Synthesis* (Savoy Place, London: UK), pp. 14/1 - 14/8.

LUMBRERAS, M., ROSSI, G., 1995, A metaphor for the visually impaired: browsing information in a 3D aural environment only, In *Proceedings of the ACM Conference on Human Factors in Computing Systems, CHI 1995* (Denver: USA), pp. 261-262.

- MacCORMAC, E.R., 1985, *A cognitive theory of metaphor* (MIT Press, Cambridge, MA).
- MacCORMAC, E.R., 1976, *Metaphor and Myth in Science and Religion* (Durham: Duke University Press).
- MACK, R.L., LEWIS, C.H., and CARROLL, J.M., 1983, Learning to use word processors: problems and prospects. *ACM Transactions in Office Information Systems*, 1, 254-271.
- MAGLIO, P.P., and MATLOCK, T., 1998, Metaphors we surf the web by, In *Proceedings of the Workshop on Personal and Social Navigation in Information Space* (Stockholm: Sweden), pp. 138-149.
- MAGUIRE, M.C., 2001a, Methods to support human-centred design. *International Journal of Human-Computer Studies*, 55(4), 587-634.
- MAGUIRE, M.C., 2001b, Context of use within usability activities. *International Journal of Human Computer Studies*, 55, 453-483.
- MALONE, T.W., 1983, How do people organise their desks? implications for the design of office information systems. *ACM Transactions in Office Information Systems*, 1(1), 99-112.
- MARCUS, A., 2002, Metaphors and user interfaces in the 21st century, *Interactions*, 9(2), pp. 7-10.
- MARICS, M.A., and ENGELBECK, G, 1997, Designing voice menu applications for telephones. In: *Handbook of Human-Computer Interaction*, edited by M.G. Helander, T.K. Landauer, and P. Prabhu (Elsevier: New York), pp. 1085-1102.
- MARKOWITZ, J., 1993, The power of speech. *AI Expert*, 29, pp. 33.

MARTIN, G.L., 1989, The Utility of speech input in user-computer interfaces. *International Journal of Man Machine Studies*, **30**(4), 355-375.

MARTIN, M.M., WILLIGES, B.H., and WILLIGES, R.C., 1990, Improving the design of telephone-based systems, In Proceedings of the Human Factors Society 34th Annual meeting, pp. 198-202.

MARTIN, T.B., 1976, Practical applications of voice input to machines. Proceedings of the IEEE, **64**(4), 487-501.

MARX, M., and SCHMANDT, C., 1996, MailCall: message presentation and navigation in a nonvisual environment, In Proceedings of the Conference on Human Factors in Computing Systems, pp. 165 - 172.

MAULSBY, D., GREENBERG, S., and MANDER, R., 1993, Prototyping an intelligent agent through Wizard of Oz, In Proceedings of Human Factors in Computing Systems, INTERCHI 1993, pp. 277-284.

MAYER, R.E., 1975, Different problem-solving competencies established in learning computer programming with and without meaningful models. *Journal of Educational Psychology*, **67**, 725-734.

MAYER, R.E., 1981, The psychology of how novices learn computer programming. *Computer Surveys*, **13**, 121-141.

MCDONALD, J.E., and SCHVANEVELDT, R.W., 1988, The application of user knowledge to interface design. In: *Cognitive Science and its Applications for Human-Computer Interaction*, edited by R. Guindon (Hillsdale: Lawrence Erlbaum), pp. 289-338.

McGEE, M.G., 1979, Human spatial abilities: psychometric studies and environmental, genetic, hormonal, and neurological influences. *Psychological Bulletin*, **86**, 889-911.

- MCINNES, F.R., NAIRN, I.A., ATTWATER, D.J., and JACK, M.A., 1999, Effects of prompt style on user responses to an automated banking service using word-spotting. *BT Technology Journal*, **17**(1), pp. 160-171.
- MCINNES, F.R., JACK, M.A., CARRARO, F., and FOSTER, J.C., 1997, User responses to prompt wording styles in a banking service with a Wizard of Oz simulation of word-spotting, In Proceedings of the IEEE Colloquium on Advances in Interactive Voice Technologies for Telecommunications Services, pp. 7/1-6.
- MESSICK, S., 1976, Personality consistencies in cognition and creativity. In: *Individuality in learning: Implications of cognitive styles and creativity for human development*, edited by S. Messick (San Francisco: Jossey-Bass).
- MICHAELIS, P.R., and WIGGINS, R.H., 1982, A human factors engineer's introduction to speech synthesizers. In: *Directions in Human-Computer Interaction*, edited by A. Badre and B. Shneiderman (Norwood, N.J: Ablex Publishing Corp.), pp. 149-178.
- MILES, M.B., and HUBERMAN, A.M., 1994, *Qualitative data analysis: an expanded sourcebook*, 2nd edition (Sage Publications, Thousand Oaks, CA).
- MIYAKE, N., 1986, Constructive interaction and the iterative process of understanding. *Cognitive Science*, **10**, 151-177.
- MOHAGEG, M.F., 1992, The influence of hypertext linking structures on the efficiency of information retrieval, *Human Factors*, **34**(3), pp. 351-367.
- MOORE, T.J., 1989, Speech technology in the cockpit. In: *Aviation Psychology*, edited by R.S. Jensen (Aldershot: Gower Technical).
- MORGAN, G., 1996, An afterword: is there anything more to be said about metaphor. In: *Metaphor and Organizations*, edited by D. Grant, and C. Oswick (Thousand Oaks, California: Sage Publications).

MORGAN, K., and MACLEOD, H., 1990, The possible role of personality factors in computer interface preference, In Proceedings of the second interdisciplinary workshop on mental models (Cambridge: UK).

MORRISON, D.L., GREEN, T.R.G., SHAW, A.C. and PAYNE, S.J., 1984, Speech-controlled text-editing: effects of input modality and command structure. *International Journal of Man - Machine Studies*, **21**, 49 - 64.

MOUNTFORD, J.S., 1995, Tools and techniques for creative design. In: *Readings in Human-Computer Interaction: Toward the Year 2000*, 2nd edition, edited by R.M. Baecker, J. Grudin, W.A.S. Buxton, and S. Greenberg (Morgan Kaufmann), pp. 128-141.

MULLET, K., and SANO, D., 1995, *Designing Visual Interfaces: Communication Oriented Techniques* (SunSoft Press, Prentice Hall).

MURDOCK, B.B., and WALKER, K.D., 1969, Modality effects in free recall. *Journal of Verbal Learning and Verbal Behaviour*, **8**, 665-676.

MURRAY, D., and BEVAN, N., 1984, The social psychology of computer conversations, In Proceedings of Human-Computer Interaction, INTERACT 1984 (London: UK).

MYNATT, E., and EDWARDS, W.K., 1992, Mapping GUIs to auditory interfaces, In Proceedings of ACM Symposium on User Interface Software and Technology (Monterey, California: United States), pp. 61-70.

MYNATT, E., and EDWARDS, W.K., 1995, Metaphors for nonvisual computing. In: *Extraordinary Human-Computer Interaction : Interfaces for Users with Disabilities*, edited by A. Edwards, and J. Long (Cambridge University Press, New York, NY, USA), pp. 201-220.

- MYNATT, E.D., 1993, Auditory presentation of graphical user interfaces. In: *Sonification, Audification and Auditory Interfaces*, edited by G. Kramer (Santa Fe, Addison-Wesley: Reading, MA).
- MYNATT, E.D., 1995, Transforming graphical interfaces into auditory interfaces, In *Proceedings of CHI 1995* (Denver, Colorado: United States), pp. 67-68.
- NARDI, B.A., and ZARMER, C.L., 1993, Beyond models and metaphors: visual formalisms in user interface design. *Journal of Visual Languages and Computing*, **4**, 5-33.
- NEALE, D.C., and CARROLL, J.M., 1997, The role of metaphors in user interface design. In: *Handbook of Human-Computer Interaction*, edited by M.G. Helander, T.K. Landauer, and P. Prabhu (Elsevier: New York), pp. 441-462.
- NEALE, H., and NICHOLS, S., 2001, Theme-based content analysis: a flexible method for virtual environment evaluation. *International Journal of Human-Computer Studies*, **55**, 167-189.
- NEISSER, U., 1967, *Cognitive Psychology* (Appleton Century Crofts, NY).
- NELSON, T., 1990, The right way to think about software design. In: *The Art of Human Computer Interface Design*, edited by B. Laurel, and S. Mountford (New York: Addison-Wesley Inc).
- NELSON, D.L., 1986, User acceptance of voice recognition in a product inspection environment, In *Proceedings of Speech Tech 1986: Voice Input/Output Applications Show and Conference* (New York: Media Dimensions Inc), pp. 62.
- NERSESSIAN, N.J., 1995, Capturing the dynamics of conceptual change in science. In: *Diagrammatic Reasoning: Cognitive and Computational Perspectives*, edited by J. Glasgow, N.H. Narayanan and B. Chandrasekaran (Menlo Park, CA: AAAI Press), pp. 137-181.

- NEWELL, A.F., 1992, Wither speech systems? Some characteristics of spoken language which may effect the commercial viability of speech technology. In: *Advances in Speech, Hearing, and Language Processing* (JAI Press Ltd, London).
- NEWELL, A.F., 1987, Speech simulation studies – performance and dialogue specification, In Proceedings of Unicom seminar, Recent Developments and Applications of Natural Language Understanding (London).
- NICKERSON, R.S., 1976, On conversational interaction with computers. In: *Readings in Human Computer Interaction*, edited by R.M. Baecker, and W.A.S. Buxton (Los Altos, CA: Morgan Kaufmann), pp. 681-693.
- NIELSEN, J., 1990, A meta-model for interacting with computers. *Interacting with Computers*, **2**, 147-160.
- NIELSEN, J., 1993, *Usability Engineering* (Cambridge, MA: Academic Press).
- NIELSEN, J., 1994, (Ed.), Usability laboratories. Special double issue of: *Behaviour and Information Technology*, **13**, 1-2.
- NOLDER, R., 1991, Mixing metaphor and mathematics in the secondary classroom. In: *Language in Mathematical Education: Research and Practice*, edited by K. Durkin and B. Shire (Buckingham, UK: Open University Press), pp. 105-114.
- NORCIO, A.F., and STANLEY, J., 1989, Adaptive human-computer interfaces: a literature survey and perspective. *IEEE Transactions on Systems, Man and Cybernetics*, **19**, 399-408.
- NORMAN, D.A., 1983, Some observations on mental models. In: *Mental Models*, edited by D. Gentner, and A. Stevens (Hillsdale, NJ: Erlbaum).
- NORMAN, D.A., 1986, Cognitive engineering. In: *User-Centred System Design: New perspectives in human-computer interaction*, edited by D.A. Norman, and S.W. Draper (Hillsdale, NJ: LEA).

NORMAN, D.A., 1988, *The psychology of everyday things* (New York: Basic Books).

NORMAN, D.A., 1993, *Things that make us smart* (Reading, MA: Addison-Wesley).

NORMAN, D.A., and DRAPER, S.W., 1986, *User centred system design: New perspectives on human-computer interaction* (Hillsdale, NJ: Erlbaum).

NOYES, J., 2001, Talking and writing-how natural in human-machine interaction? *International Journal of Human-Computer Studies*, **55**(4), pp. 503-519

NOYES, J.M., and FRANKISH, C.F., 1989, A review of speech recognition applications in the office. *Behaviour and Information Technology*, **8**(6), 475-86.

NOYES, J.M., HAIGH, R., and STARR, A.F., 1989, Automatic speech recognition for disabled people. *Applied Ergonomics*, **20**(4), 293-298.

NYE, J.M., 1982, Human factors analysis of speech recognition systems, *Speech Technology*, **1**, 50-57.

OLTMAN, P.K., RASKIN, E., and WITKIN, H.A., 1971, *Group embedded figures test* (Palo Alto, CA: Consulting Psychologists Press).

OREN, T., SALOMON, G., KREITMAN, K., and DON, A., 1990, Guides: characterizing the interface. In: *The art of human computer interface design*, edited by B. Laurel (Addison-Wesley, Reading, Mass), pp. 367-381.

ORTONY, A., 1975, Why metaphors are necessary and not just nice. *Educational Theory*, **25**(1), 45-53.

OSBORN, A.F., 1963, *Applied Imagination* (New York: Schribeners and Sons).

- PAAP, K.R., COOKE, N.J., 1997, Design of menus. In: *Handbook of human-computer interaction*, 2nd edition, edited by M. Helander, T.K. Landauer, and P. Prabhu (Amsterdam: Elsevier Science), pp. 533-572.
- PALEN, L., SALZMAN, M., and YOUNGS, E., 2001, Discovery and integration of mobile communications in everyday life. *Personal and Ubiquitous Computing Journal*, **5**, pp. 109-122.
- PALMQUIST, R.A., 2001, Cognitive style and users metaphors for the web: an exploratory study, *The Journal of Academic Librarianship*, **27**(1), 24-32.
- PASK, G., 1988, Learning strategies, teaching strategies, and conceptual or learning style. In: *Learning Strategies and Learning Styles*, edited by R.R. Schmeck (New York: Plenum Press), 83-100.
- PC ADVISOR, 2003, Mobile users to reach 1 billion, Issue 91, Spring, URL: <http://www.pcadvisor.co.uk>
- PEACOCK, G.E., 1984, Humanising the man-machine interface, *Speech Technology*, **2**, 106-108.
- PECKHAM, J., 1993, A new generation of spoken dialogue systems: results and lessons from the SUNDIAL project, In Proceedings of Eurospeech 1993, pp. 33-40.
- PELLEGRINO, J.W., 1985. Anatomy of analogy. *Psychology Today*, 49-54.
- PELLEGRINO, J.W., ALDERTON, D.L., and SHUTE, V.J., 1983, Understanding spatial ability. *Educational Psychologist*, **19**, 239-253.
- PIRHONEN, A., and BREWSTER, S., 2001, Metaphors and imitation. In Proceedings of the PC-HCI Workshop 2001, (Patras: Greece), pp. 27-32.
- PLATO, 1955, *The Republic* (London: Penguin). (Original work written ca. 380BC).

POOCK, G.K., MARTISA, B.J. and ROLAND, E.F., 1983, The effects of feedback to users with voice recognition equipment. *Report No. NP55-83-003* (Monteray, CA: Naval Postgraduate School).

POPPER, K., 1958, *The Logic of Scientific Enquiry* (London: Hutchinson).

POTOSNAK, K., 1988, Do icons make user interfaces easier to use? *IEEE Software*, pp. 97-99.

POULSON, D.F., 1987, Towards simple indices of the perceived quality of software interfaces, In *Proceedings of the IEE Colloquium on Evaluation Techniques for Interactive Systems Design* (London: Institute of Electrical Engineers).

PREECE, J.J., ROGERS, Y.R., SHARP, H., BENYON, D.R., HOLLAND, S., and CAREY, T., 1994, *Human-Computer Interaction* (Addison-Wesley).

PREECE, J.J., ROGERS, Y.R., SHARP, H., 2002, *Interaction Design: Beyond Human-Computer Interaction* (JohnWiley and Sons).

PUGH, S., HICKS, J., and DAVIS, M., 1997, Metaphorical ways of knowing: the imaginative nature of thought and language (Urbana, IL: NCTE).

QIN, Y., and SIMON, H., 1995, Imagery and mental models. In: *Diagrammatic Reasoning: Cognitive and Computational Perspectives*, edited by J. Glasgow, N.H. Narayanan, and B. Chandrasekaran (Menlo Park, CA: AAAI Press), pp. 403-434.

RABINER, L.R., 1995, Applications of voice processing to telecommunications, *Proceedings of IEEE*, **82**(2), 199-230.

REDDY, M.J., 1979, The conduit metaphor: a case of frame conflict in our language about language. In: *Metaphor and Thought*, edited by A. Ortony (London: Cambridge University Press), pp. 284-324.

REEVES, B., and NASS, C., 1999, *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places* (Cambridge University Press, New York, NY).

RHYNE, J.R. and WOLF, C.G., 1993, Recognition-based user interfaces. In: *Advances in Human-Computer Interaction*, edited by H.R. Hartson and D. Hix (Ablex Publishing Corp: Norwood, N.J.), pp. 191-250.

RICHARDS, I.A., 1936, *The Philosophy of Rhetoric* (Oxford: Oxford University Press).

RICHARDS, M.A., and UNDERWOOD, K., 1984, Talking to machines: How are people naturally inclined to speak. In: *Contemporary Ergonomics*, edited by E.D. Megaw (London: Taylor and Francis).

RIDING, R.J., and SADLER-SMITH, E., 1992, Type of instructional material, cognitive style and learning performance. *Educational Studies*, **18**, pp. 323- 340.

RIDING, R.J., 1991, *Cognitive Styles Analysis users' manual* (Birmingham, Learning & Training Technology).

ROBERTSON, S., 2001, Requirements trawling: techniques for discovering requirements. *International Journal of Human-Computer Studies*, **55**, 405-421.

ROBSON, C., 2002, *Real World Research* (Oxford: Blackwell Publishing).

ROGERS, Y., LEISER, R., and CARR, D., 1988, Evaluating metaphors for use at the user-system interface. In *Proceedings of the first European Conference on Information Technology for Organisational Systems* (Elsevier: Amsterdam).

ROLLINS, A., CONSTANTINE, B., and BAKER, S., 1983, Speech recognition at two field sites, In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (Boston, Massachusetts: United States), pp. 267-273.

- ROSE, S., 1993, *The Making of Memory* (New York: Bantam Books).
- ROSENFELD, L., and MORVILLE, P., 1998, *Information Architecture for the World Wide Web* (Cambridge: O'Reilly).
- ROSENTHAL, R., and ROSNOW, R.L., 1975, *The volunteer subject* (New York: Wiley).
- ROSSON, M.B., 1983, Patterns of experience in text editing, In Proceedings of CHI 1983 (Boston, Massachusetts, United States), pp. 171–175.
- ROSSON, M.B., 1985, Using synthetic speech for remote access to information. *Behaviour, Research Methods, Instruments and Computers*, 17(2), 250-252.
- ROSSON, M.B., and CECALA, A.J., 1986, Designing a quality voice: an analysis of listeners' reactions to synthetic voices, In Proceedings of CHI 1986, pp. 192-197.
- RUMELHART, D., and NORMAN, D., 1981, Analogical processes in learning. In: *Cognitive Skills and their Acquisition*, edited by J.R. Anderson (Erlbaum Associates: Hillsdale, NJ), pp. 335-361.
- RUST, J., and GOLOMBOK, S., 1989, *Modern Psychometrics: The science of psychological assessment* (Routledge: London and New York).
- RUTTER, D.R., 1987, *Communication by telephone* (Oxford: Pergamon Press).
- SACKS, H., SCHLEGOFF, E., and JEFFERSON, G., 1974, A simple systematics for the organisation of turn taking for conversation. In: *Studies in the organisation of conversational interaction*, edited by J. Schenkin (NY: Academic Press).
- SALTHOUSE, T.A., BABCOCK, R.L., SKOVRONEK, E., MITCHELL, D.R., and PALMON, R., 1990, Age and experience effects in spatial visualization. *Developmental Psychology*, 26, 128-136.

- SALVENDY, G., and KNIGHT, J., 1982, Psychomotor work capabilities. In: *Handbook of industrial engineering*, edited by G. Salvendy (New York: Wiley).
- SAVIDIS, A., and STEPHANIDIS, C., 1995, Developing dual user interfaces for integrating blind and sighted users: The HOMER UIMS, In Proceedings of the ACM Conference on Human Factors in Computing Systems, CHI 1995 (Denver: United States), pp. 106-113.
- SAWHNEY, N., and SCHMANDT, C., 1998, Speaking and listening on the run: design for wearable audio computing, In Proceedings of the Second International Symposium on Wearable Computers (ISWC), pp. 108-115.
- SCHANK, R., and ABELSON, R., 1977, *Scripts Plans Goals and Understanding: An Inquiry into Human Knowledge Structures* (New Jersey: LEA Publishers).
- SCHILIT, B., and THEIMER, M., 1994, Disseminating active map information to mobile hosts. *IEEE Network*, 8(5), 22-32.
- SCHMANDT, C., 1994, *Voice communication with computers: Conversational systems* (New York: Van Nostrand Reinhold).
- SCHMANDT, C., and ARONS, B., 1984, Phoneslave: a graphical telecommunications system, In Proceedings of the Society for Information Display International Symposium (San Francisco, CA: United States), pp. 146-149.
- SCHON, D.A., 1979, Generative metaphor and social policy. In: *Metaphor and Thought*, edited by A. Ortony (Cambridge: Cambridge University Press), pp. 137-163.
- SCHUMACHER, R.M., 1992, Phone-based interfaces: Research and guidelines. In Proceedings of the Human Factors Society 36th Annual Meeting (Santa Monica, CA: Human Factors Society), pp. 1051-1055.
- SCHUMACHER, R.M., HARDZINSKI, M., and SCHWARTZ, A., 1995, Increasing the usability of interactive voice response systems. *Human Factors*, 37(2), 251-264.

SEIN, M. K., and BOSTROM, R. P., 1989, Individual differences and conceptual models in training novice users. *Human-Computer Interaction*, 4, 197-229.

SEIN, M.K., OLFMAN, L., BOSTROM, R.P., and DAVIS, S.A., 1993, Visualization ability as a predictor of user learning success. *International Journal of Man-Machine Studies*, 39, 599-620.

SHACKEL, B., 1981, The concept of usability, In Proceedings of IBM Software and Information Usability Symposium (Poughkeepsie, NY: United States), pp. 1-30.

SHEPARD, R.N., 1978, Externalization of mental images and the act of creation. In: *Visual Learning, Thinking and Communication*, edited by B.S. Randhawa and W.E.Coffman (New York, Academic Press), pp. 133-189.

SHEPHERD, A., 1989, Analysis and training in information technology tasks. In: *Task Analysis for Human-Computer Interaction*, edited by D. Diaper (Chichester, Ellis Horwood).

SHERIDAN, T.B., CHARNEY, L., MENDEL, M.B., and ROSEBOROUGH, J.B., 1986, Supervisory control, mental models, and decision aids. *MIT Department of mechanical Engineering technical report* (Massachusetts Institute of Technology, Cambridge, MA).

SHIH, C., and GAMON, J., 1999, Student learning styles, motivation, learning strategies, and achievement in web-based courses, Available: <http://iccel.wfu.edu/publications/journals/jcel/jcel1990305/ccshih.htm>.

SHNEIDERMAN, B., 1987, *Designing the user interface: Strategies for effective human-computer interaction* (Reading, MA: Addison-Wesley).

SHNEIDERMAN, B., 1992, *Designing the user interface: Strategies for effective human-computer interaction*, 2nd edition (Reading, MA: Addison-Wesley).

SHNEIDERMAN, B., 1998, *Designing the user interface: Strategies for effective human-computer interaction*, 3rd edition, (Reading, MA: Addison-Wesley).

SLOWIACZEK, L.M., and NUSBAUM, H.C., 1985, Effects of speech rate and pitch contour on the perception of synthetic speech. *Human Factors*, **27**(6), 701-712.

SMILOWITZ, E.D., 1996, Do metaphors make Web browsers easier to use? In Proceedings of the 2nd Microsoft Conference in Human Factors and the Web. Designing for the Web: Empirical Studies. Available at: <http://www.baddesigns.com/mswebcnf.htm>

SMITH, D, C., IRBY, C., KIMBALL, R., VERPLANK, B. and HARSLEM, E. 1982, Designing the Star user interface. *Byte*, **7**(4), 242-282.

SMYTH, M., ANDERSON, B., and ALTY, J.L., 1995, Metaphor reflections and a tool for thought, In Proceedings of HCI Conference 1995, People and Computers X (Huddersfield: UK), pp. 137-150.

SMYTH, M., and KNOTT, R.P., 1994, The role of metaphor at the human computer interface, In Proceedings of OZCHI 1994, Harmony through working together (Melbourne: Australia), pp. 287-291.

SOUVIGNIER, V., KELLNER, A., RUEBER, B., SCHRAMM, H., and SEIDE, F., 2000, The thoughtful elephant: strategies for spoken dialogue systems, *IEEE Transactions of Speech Audio Process*, **8**(1), 51-62.

SPIRO, R.J., FELTOVICH, P.J., COULSON, R.L., and ANDERSON, D.K., 1989, Multiple analogies for analogy-induced misconception in advanced knowledge acquisition. In: *Similarity and Analogical Reasoning*, edited by S. Vosniadou, and A. Ortony (Cambridge, England: Cambridge University Press), pp. 498-531.

STAGGERS, N., and NORCIO, A.F., 1993, Mental models: concepts for human-computer interaction research. *International Journal of Man-Machine Studies*, **38**, 587-605.

- STANNEY, K., and SALVENDY, G., 1995, Information Visualisation: Assisting low spatial individuals with information access tasks through the use of visual mediators. *Ergonomics*, **38**(6), 1184-1198.
- STENNING, K., and OBERLANDER, J., 1995, A cognitive theory of graphical and linguistic reasoning: logic and implementation, *Cognitive Science*, **19**, 97-140.
- STENTIFORD, F.W.M., and POPAY, P.A., 1999, The design and evaluation of dialogues for interactive voice response services, *BT Technology Journal*, **17**(1), 142-148.
- STRATHMEYER, C.R., 1990, Voice in computing: an overview of available technologies, *IEEE computer Magazine*, **23**(8), 10-15.
- STUMPF, H., and ELIOT, J., 1995, Gender-related differences in spatial ability and the k factor of general spatial ability in a population of academically talented students. *Personality and Individual Differences*, **19**, 33-45.
- TATE, M., WEBSTER, R., and WEEKS, R., 1993, Evaluation and prototyping of dialogues for voice applications. In: *Interactive Speech Technology*, edited by C. Baber and J.M. Noyes (London: Taylor and Francis), pp. 157-165.
- THE BRITISH PSYCHOLOGICAL SOCIETY, 1993, Code of Conduct, Ethical Principles and Guidelines.
- THOMAS, C., and BEVAN, N., 1995, *Usability context analysis: a practical guide* (National Physical Laboratory, Teddington: UK).
- THORNDYKE, P.W., and STASZ, C., 1980, Individual differences in procedures for knowledge acquisition from maps. *Cognitive Psychology*, **12**, 137-175.
- TREISMAN, A., and DAVIES, A., 1973, *Divided attention to eye and ear* (New York: Academic Press).

TUN, P.A., and WINGFIELD, A., 1997, Language and communication: fundamentals of speech communication and language processing in old age. In: *Handbook of human factors and the older adult*, edited by A.D. Fisk, and W.A. Rogers (Academic Press, San Diego), pp. 125-149.

TURNER, M., 1987, *Death is the Mother of Beauty: Mind, Metaphor, Criticism* (Chicago: University of Chicago Press).

VAN DER VEER, G.C., 1990, Human-Computer Interaction: Learning, Individual Differences and Design Recommendations. *PhD Thesis* (Free University of Amsterdam).

VENKATESH, V., and MORRIS, M.G., 2000, Why don't men ever stop to ask for directions? Gender, social influence, and their role in technology acceptance and usage behaviour, *MIS Quarterly*, **24**, 115-139.

VERPLANK, B., 1990, Graphic invention for user-interface design, In Proceedings of the CHI 1990 Workshop (Seattle, Washington: United States).

VETERE, F., and HOWARD, S., 1999, Redundancy and prior knowledge, In Proceedings of OZCHI 1999, Conference of the Computer-Human Interaction Special Interest Group of the Ergonomics Society of Australia (Wagga Wagga: Australia).

VICENTE, K.J., HAYES, B.C., and WILLIGES, R.C., 1987, Assaying and isolating individual difference in searching a hierarchical file system. *Human Factors*, **29**, 349-359.

VICENTE, K., and WILLIGES, R., 1988, Accommodating individual differences in searching a hierarchical file system. *International Journal of Man-Machine Studies*, **29**, 647-668 .

VISICK, D., JOHNSON, P., and LONG, J., 1984, The use of simple speech recognisers in industrial applications, In Proceedings of INTERACT 1984, the First

IFIP conference on Human-Computer Interaction (London: International Federation of Information Processing Societies).

WARE, C., 2000, *Information Visualization: Perception for design* (San Francisco, CA, Morgan Kaufmann).

WARE, C., and OSBORNE, S., 1990, Exploration and virtual camera control in virtual three dimensional environments, In Proceedings of the 1990 Symposium on Interactive 3D Graphics, Special Issue of Computer Graphics, pp. 175-183.

WECHSLER, D., 1952, *The range of human capabilities*, 2nd edition (Baltimore, MD: Williams and Wilkins).

WEI, R., and LEUNG, L., 1999, Blurring public and private behaviours in public space: policy challenges in the use and improper use of the cell phone. *Telematics and Informatics*, **16**, 11-26.

WELLER, H.G., REPMAN, J., and ROOZE, G.E., 1994, The relationship of learning, behavior, and cognitive styles in hypermedia-based instruction: Implications for design of HBI. *Computers in the Schools*, **10**, 401-420.

WESTALL, F.A., JOHNSON, R.D., and LEWIS, A.V., 1996, Speech technology for telecommunications. *BT technology Journal*, **14**(1), 9-27.

WHITESIDE, J., BENNETT, J., and HOLTZBLATT, K., 1988, Usability engineering: our experience and evolution. In: *Handbook of Human-Computer Interaction*, edited by M. Helander (Amsterdam: Elsevier), pp. 791-817.

WHITLEY, B.E., 1997, Gender differences in computer-related attitudes and behaviours: A meta-analysis. *Computers in Human Behaviour*, **13**, 1-22.

WHITTAKER, S.J., and ATTWATER, D.J., 1996, Interactive speech systems for telecommunications applications. *BT Technology Journal*, **14**(2), 11-23.

- WICKENS, C.D., MOUNTFORD, S.J. and SCHREINER, W., 1981, Multiple resources task, hemispheric integrity, and individual differences in time sharing. *Human Factors*, **23**, 211-229.
- WICKENS, C.D., 1992, *Engineering Psychology and Human Performance*, 2nd edition (Harper Collins, NY).
- WILKINSON, S., CRERAR, A., and FALCHIKOV, N., 1997, Book versus hypertext: Exploring the association between usability and cognitive style. Available: <http://www.dcs.napier.ac.uk/~simon/results/abstract.htma>
- WILLIAMS, M.D., HOLLAN, J.D., and STEVENS, A.L., 1983, Human reasoning about a simple physical system. In: *Mental models*, edited by D. Gentner, and A.L. Stevens (Lawrence Erlbaum Associates, Hillsdale, NJ), pp. 134-154.
- WILPON, J.G., and ROBERTS, L.A., 1986, The effects of instructions and feedback on speaker consistency for automatic speech recognition, In Proceedings of the IEEE International Conference on Speech Input/Output (London: England), pp. 242-247.
- WITKIN, H.A., MOORE, C.A., GOODENOUGH, D.R., and COX, P.W., 1977, Field-dependent and field independent cognitive styles and their educational implications. *Review of Educational Research*, **47**, 1-64.
- WIXON, D., and WILSON, C., 1997, The usability engineering framework for product design and evaluation. In: *Handbook of human-computer interaction*, edited by M.G. Helander, T.K. Landauer, and P. Prabhu, 2nd edition (Amsterdam, The Netherlands: North-Holland).
- WOLF, C.G., KASSLER, M., ZADROZNY, W., and OPYRCHAL, L., 1997, Talking to the conversation machine: an empirical study, In Proceedings of the IFIP TC13 International Conference on Human-Computer Interaction, pp. 461-468.
- WOLF, C., KOVED, L., and KUNZINGER, E., 1995, Ubiquitous Mail: Speech and graphical interfaces to an integrated voice/email mailbox, In Proceedings of IFIP

International Conference on Human Computer Interaction, Interact 1995, Chapman & Hall, pp. 247-252.

YANKELOVICH, N., LEVOW, G. and MARX, M., 1995, Designing SpeechActs: Issues in speech user interfaces, In Proceedings of CHI 1995, ACM, New York, pp. 369-376.

YANKELOVICH, N., 1994, Talking vs taking: speech access to remote computers, In Proceedings of the 1994 Conference on Human Factors in computing systems (Boston, Massachusetts, United States), pp.275-276.

YOUNG, R.M., 1983, Surrogates and mappings: Two kinds of conceptual models for interactive devices. In: *Mental models*, edited by A.L. Stevens, and D. Gentner (Hillsdale, NJ: Lawrence Erlbaum Associates), pp. 35-52.

ZAJICEK, M.P., 1990, Evaluation of a speech driven interface, In Proceedings of the UK IT 1990 Conference (Southampton: UK), pp. 286-293.

ZANDRI, E., and CHARNESS, N., 1989, Training older and younger adults to use software. *Educational Gerontology*, **15**, 615-631.

ZOLTAN-FORD, E., 1991, How to get people to say and type what computers can understand. *International Journal of Man-Machine Studies*, **34**, 527-547.

ZUE, V., and GLASS, J., 2000, Conversational interfaces: advances and challenges, *Proceedings of the IEEE*, **88**(8), 1166-1180.

:: APPENDIX 1

Likert questionnaire statements matched to the 6 subjective factors

Subjective factor from SASSI	Statement from the Likert questionnaire
System response accuracy	The service was efficient
	The service was reliable
	The service was unpredictable
	The service didn't always do what I wanted
	The service didn't always do what I expected
	The service was dependable
	The service made few errors
	The service was consistent
Likeability	The service was useful
	The service was pleasant
	I was able to recover easily from errors
	It was clear how to speak to the service
	The service was too polite
	The service voice was clear
	I think the service is a good idea
	I enjoyed using the service
	I felt the help offered by the service was helpful
	I would be happy using the service in future
	The service was too complicated
	The service was friendly
	I liked the service voice
	I would have preferred to speak to a human being when using the service
	I felt that the service needed a lot of improvement
Cognitive demand	I felt confident using the service
	I felt calm using the service
	I found the service easy to use
	I had to concentrate hard when using the service
	I felt flustered when using the service
	I felt I was in control while using the service
	I found it was easy to make my responses using speech
	I found the service was confusing to use

	I felt that sometimes I was given too many possibilities to choose from
	Learning to use the service was easy
Annoyance	The interaction with the service was boring
	The interaction with the service was irritating
	The interaction with the service was frustrating
	The service was too inflexible
	I felt impatient when I was using the service
	I found the wording of the messages was too repetitive
Habitability	I was not always sure what the service was doing
	It was not always obvious how to find what I wanted in the service
	The way that the choices were presented was clear
	Sometimes I felt lost while I was using the service
	When using the service I knew what I was expected to do
	It would have been useful to be able to request more help from the service
	It was not always clear what I had to say
Speed	The interaction with the service was fast
	The service responded too slowly
	I thought that it took too long to complete a task
	I found that the pace of the service was too slow

:: APPENDIX 2

Likert style usability questionnaire used for experiments one, two, and three

Please read each statement carefully, and then using the 7-point rating scale below, tick the box that best represents your agreement or disagreement with the statement.

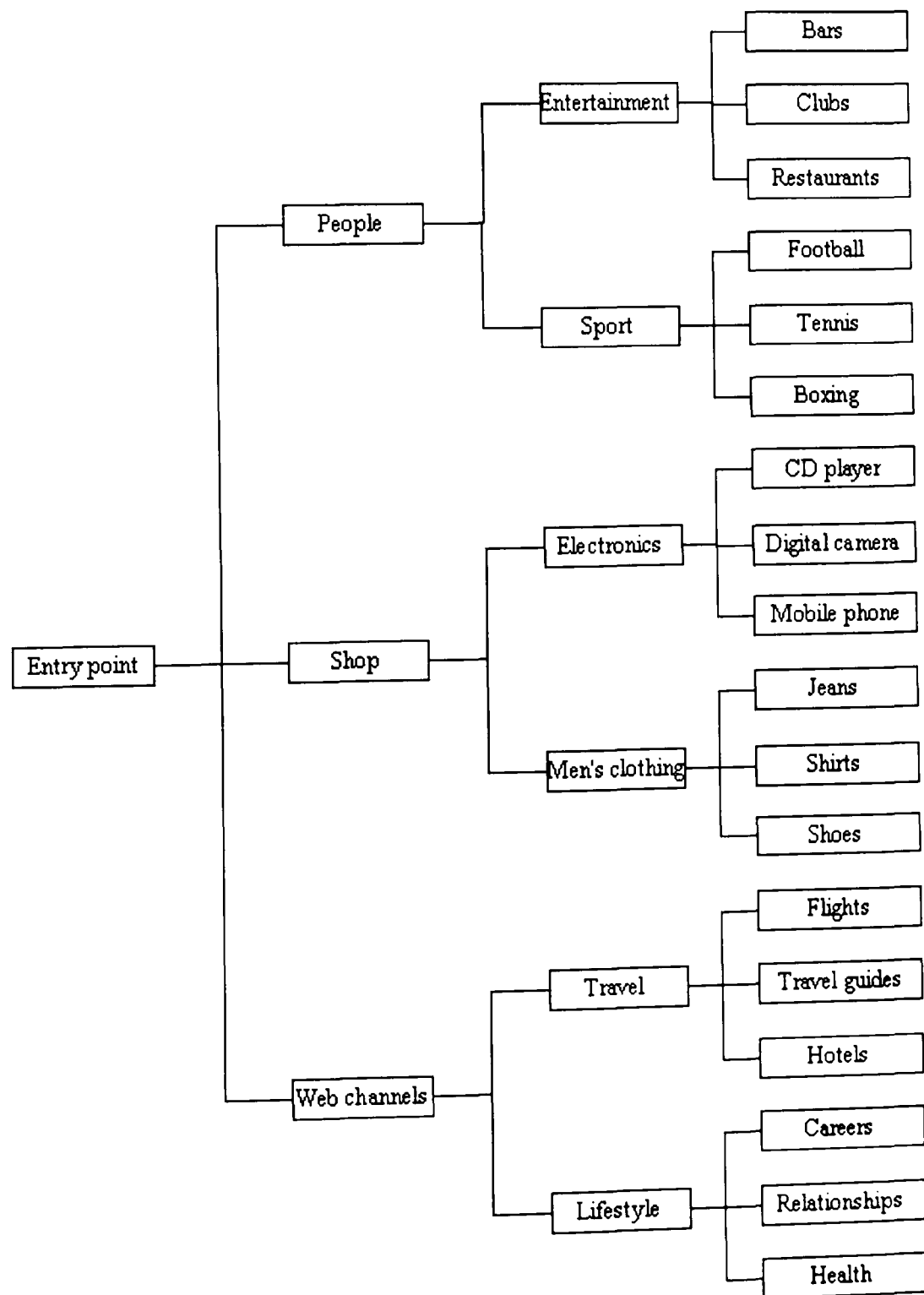
Strongly agree	Agree	Slightly agree	Neutral	Slightly disagree	Disagree	Strongly disagree
1	2	3	4	5	6	7

No	Statement	1	2	3	4	5	6	7
1	I found the service easy to use							
2	It was clear how to speak to the service							
3	The service was friendly							
4	The service was unpredictable							
5	The interaction with the service was boring							
6	The way that the choices were presented was clear							
7	Learning to use the service was easy							
8	The service was too inflexible							
9	The service responded too slowly							
10	The service was consistent							
11	I felt the help offered by the service was helpful							
12	I was not always sure what the service was doing							
13	The service was useful							
14	The service was dependable							
15	The interaction with the service was irritating							
16	I felt I was in control while using the service							
17	I would be happy using the service in future							
18	I felt that sometimes I was given too many possibilities to choose from							

19	The service voice was clear								
20	I found that the pace of the service was too slow								
21	The service didn't always do what I expected								
22	I felt confident using the service								
23	The service was efficient								
24	It was not always obvious how to find what I wanted in the service								
25	It would have been useful to be able to request more help from the service								
26	When using the service I knew what I was expected to do								
27	The service was pleasant								
28	I felt flustered when using the service								
29	Sometimes I felt lost while I was using the service								
30	I thought that it took too long to complete a task								
31	The service was too polite								
32	I felt that the service needed a lot of improvement								
33	The interaction with the service was frustrating								
34	The service made few errors								
35	I liked the service voice								
36	The service was reliable								
37	I found the wording of the messages was too repetitive								
38	The service didn't always do what I wanted								
39	I would have preferred to speak to a human being when using the service								
40	I found the service was confusing to use								
41	It was not always clear what I had to say								
42	The service was too complicated								
43	I had to concentrate hard when using the service								
44	I was able to recover easily from errors								
45	I found it was easy to make my responses using speech								
46	I felt impatient when I was using the service								
47	The interaction with the service was fast								
48	I enjoyed using the service								
49	I felt calm using the service								
50	I think the service is a good idea								

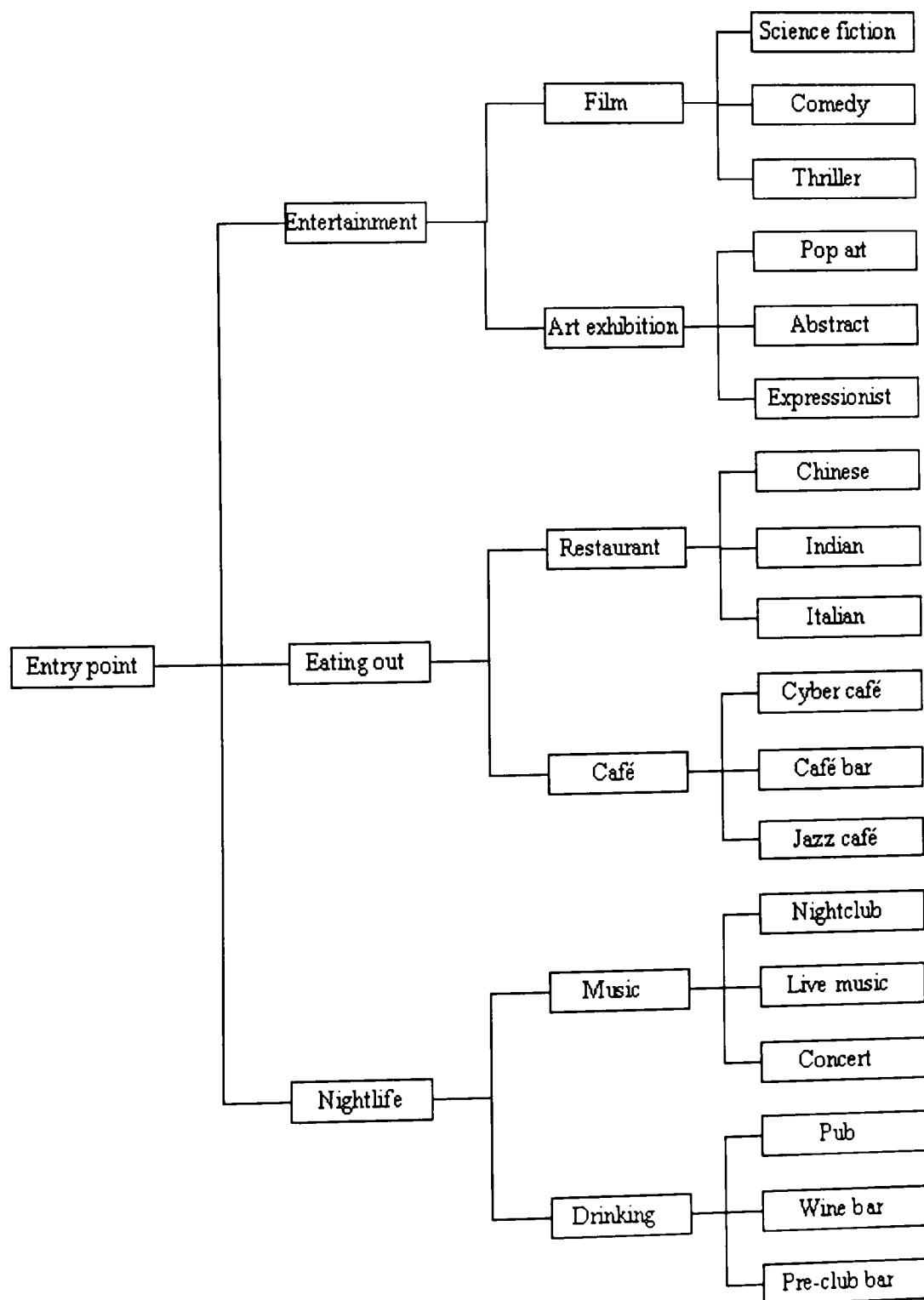
:: APPENDIX 3

The telephone Internet service structure



:: APPENDIX 4

The telephone city guide service structure



:: APPENDIX 5

The 60 'real world systems' generated in the brainstorming session

System	Ranking of 1 – 5	System	Ranking of 1 – 5
Transport		Department store	
Party		The brain	
Music		Circulatory system	
Countries/geography		University	
Cooking		Health service	
Religion		Government	
Photography		Police force	
Moving house		Computer	
Getting dressed		TV	
Fashion		Theme park	
The press		Motorway/road system	
Video recorder		Garden	
Water wheel/		Talking pages	
Car engine		Bus route	
Blind date/game show		Filing cabinet	
Bookcase		Reference book	
Crossword puzzle		Supermarket	
Drum kit		Wilderness	
CD player		Desert	
Children's playground		Night sky	
Roundabout		Argument	
Eye mechanism		Ocean	
Doctor's surgery		Museum	
House		Breakfast bar	
Room		Mandala	
Office		Sand painting	
Scaffolding		Rock garden	
Conversation		Atom/molecule	
Shopping		Cube	
Butler		Universe	

Enter your own systems in the space provided below (optional)

:: APPENDIX 6

The POPITS sheet

Please enter the name of the system you ranked as number 1 in the box below:

Which features of the system you have entered above do you consider useful in helping to explain the automated telephone service? Please complete as many of the 6 categories below as you can:

1. Physical properties of the system:

2. Operations that can be performed using the system:

3. Phrases associated with the system:

4. Images associated with the system:

5. Types of the system:

6. Sounds associated with the system:

:: APPENDIX 7

The Likert style usability questionnaire used for preliminary study one

The following statements relate to the usability of the card sorting technique. Please think carefully about each statement, and then mark your response by crossing the appropriate circle.

1. I am able to efficiently complete my work using this technique.

1 2 3 4 5 6 7
 Strongly disagree Strongly agree

2. Overall, I am satisfied with how easy it is to use this technique.

1 2 3 4 5 6 7
 Strongly disagree Strongly agree

3. It was easy to learn to use this technique.

1 2 3 4 5 6 7
 Strongly disagree Strongly agree

4. I feel comfortable using this technique.

1 2 3 4 5 6 7
 Strongly disagree Strongly agree

5. I believe I became productive quickly using this technique.

1 2 3 4 5 6 7
 Strongly disagree Strongly agree

:: APPENDIX 8

The metaphor list and ranking scores from preliminary study one

Ranking scores for the telephone Internet service

	System	Rank order of preference					Total points
		1st	2nd	3rd	4th	5th	
1	The press	1					5
2	Reference book		1			1	5
3	Conversation			1	1		5
4	Moving house				1	1	3
5	Transport			1	2	1	8
6	Computer	3	1		2		23
7	Talking pages	2	2	2			24
8	Filing cabinet	3	2	2		2	31
9	Motorway/road system	2	2		1	1	21
10	Government					1	1
11	Shopping	1	2	2		1	20
12	Fashion				1		2
13	Department store	2	1	3	1	1	26
14	Supermarket		1	3	1	2	17
15	TV		2			1	9
16	Theme park			1	1		5
17	University	1			1		7
18	Music			1			3
19	Tree		1		2		8
20	Bookcase					1	1
21	Shopping centre	1	1				9
22	Circulatory system	1		1			8
23	The brain		1			1	5
24	House		1				4
25	Bus route				1		2
26	Internet					1	1
27	Drum kit				1		2
28	Museum					1	1
29	Universe			1			3
30	Countries/geography				2		4
31	Fashion					1	1

32	Airport departure lounge	1	5
33	Breakfast bar		1

Ranking scores for the telephone city guide service

	System	Rank order of preference					Total points
		1st	2nd	3rd	4th	5th	
1	Department store	2	1	1	1		19
2	Transport	1	2			1	14
3	University		1	1			7
4	Theme park				1		2
5	Supermarket		1	1	1	3	12
6	Computer	1	2	2		1	20
7	Motorway/road system	2		1	1	1	16
8	Talking pages	2	4				26
9	Filing cabinet	1	2	1		2	18
10	Music		1		2		8
11	Party			2			6
12	Museum				2	1	5
13	Scaffolding				1		2
14	Shopping			1		1	4
15	The brain	3		1		1	19
16	Bus route		1		1	1	7
17	TV	1					5
18	Video recorder		1				4
19	CD player			1	1		5
20	Reference book				1	1	3
21	Office					1	1
22	Circulatory system		1	1	1		9
23	Tree	1	1	1			12
24	Wilderness					1	1
25	Internet	1					5
26	Universe			2			6
27	Countries/geography			1	1		5
28	Drum kit				1		2
29	Conversation			1		2	5
30	Shopping mall				1		2
31	The press	1					5
32	Entertainment guide					1	1
33	Slide viewer	1					5
34	Water wheel				1		2
35	Atom/molecule					1	1
36	Website	1					5
37	Crossword puzzle				1		2
38	Maze					1	1

:: APPENDIX 9

The Likert style usability questionnaire used for preliminary study two

Please read each statement carefully, and then using the 7-point rating scale below, tick the box that best represents your agreement or disagreement with the statement.

Strongly agree	Agree	Slightly agree	Neutral	Slightly disagree	Disagree	Strongly disagree
1	2	3	4	5	6	7

No	Statement	1	2	3	4	5	6	7
1	I found the service easy to use							
2	I found that the pace of the service was too slow							
3	I thought the service was efficient							
4	I had to concentrate hard when using the service							
5	I felt that the service needs a lot of improvement							
6	I felt flustered when using the service							
7	I felt hurried when using the service							
8	I thought that the service was reliable							
9	It would have been useful if you could have requested more help from the service							
10	I felt I was in control while using the service							
11	I would prefer to speak to a human being when using a service							
12	I found the wording of the messages too repetitive							
13	I thought the service was too polite							
14	I thought the service was friendly							
15	I liked the voices							
16	I felt impatient when I was using the service							

17	I found it was easy to make my responses using speech								
18	I thought the service was too complicated								
19	It was not always obvious how to find what I wanted in the service								
20	I thought the way that the choices were presented was clear								
21	I felt the help offered by the service was helpful								
22	I felt the service was trustworthy								
23	I would be happy using the service in future								
24	I thought that it took too long to complete a task								
25	I thought that learning to use the service was easy								
26	I enjoyed using the service								
27	I felt under stress while using the service								
28	I think the service is a good idea								
29	Sometimes I felt lost while I was using the service								
30	I thought the voices were very clear								
31	It was always clear what I had to say								
32	I found the service was confusing to use								
33	When using the service I knew what I was expected to do								
34	I felt that sometimes I was given too many possibilities to choose from								

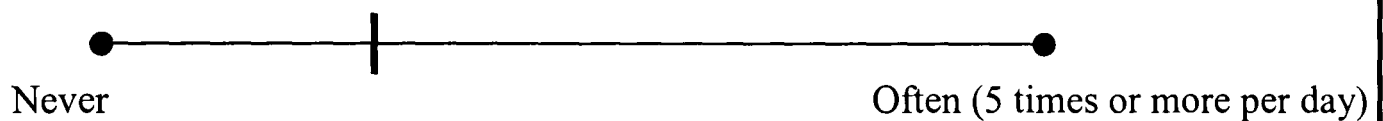
:: APPENDIX 10

The Technographic questionnaire

Please think carefully about the following questions, and then mark your response with a vertical line. Please feel free to use the full range of the scale:

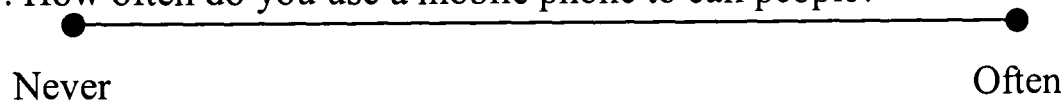
Example Question (do not attempt this question):

How often do you watch TV?

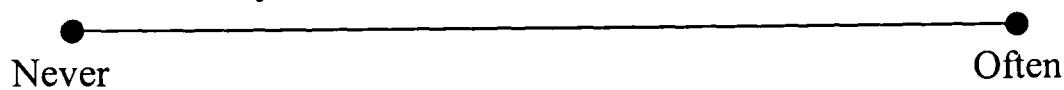


Mobile phone section:

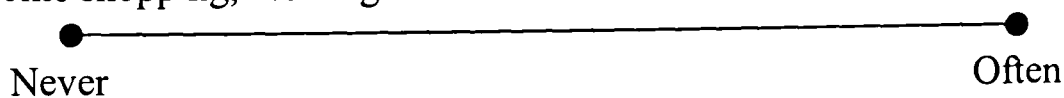
1. How often do you use a mobile phone to call people?



2. How often do you use a mobile phone to listen to voice messages?



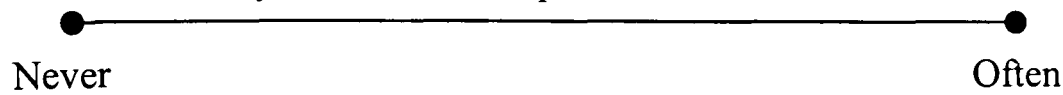
3. How often do you use a mobile phone to access automated phone services, such as home shopping, banking or cinema information?



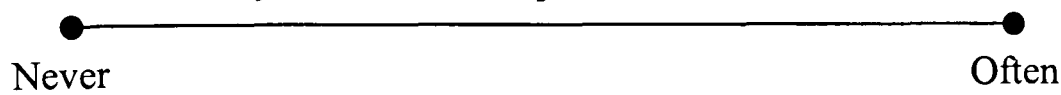
4. How often do you use speech rather than the keypad to interact with automated phone services?



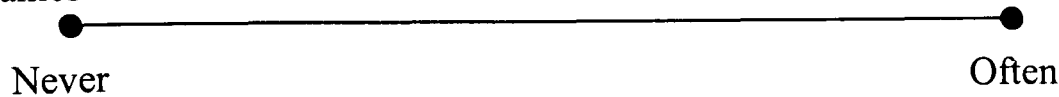
5. How often do you use a mobile phone to send or receive text messages?



6. How often do you use a mobile phone to access the Internet?

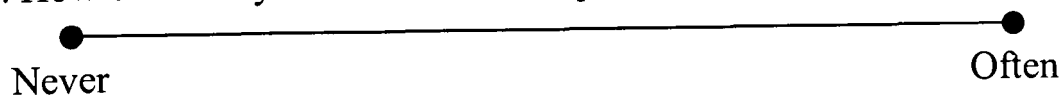


7. How often do you use a mobile phone to access services such as news and online games?

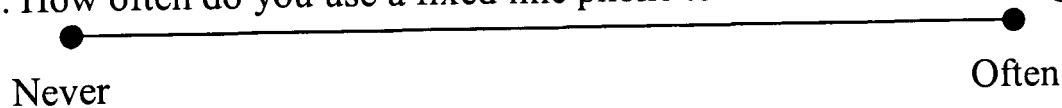


Fixed line telephone section:

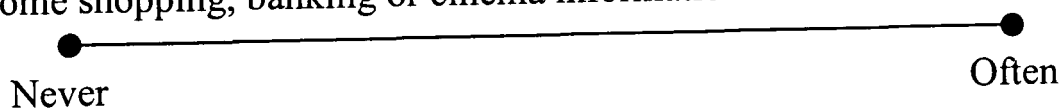
8. How often do you use a fixed line phone to call people?



9. How often do you use a fixed line phone to listen to voice messages?



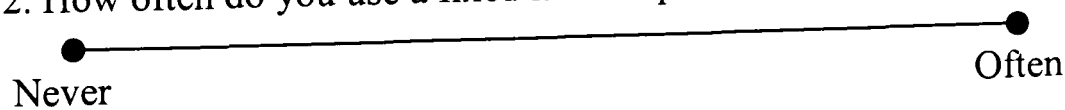
10. How often do you use a fixed line telephone to access automated services, such as home shopping, banking or cinema information?



11. How often do you use speech rather than the keypad to interact with automated phone services?

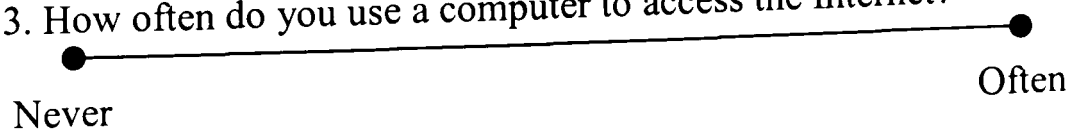


12. How often do you use a fixed line telephone to send faxes?

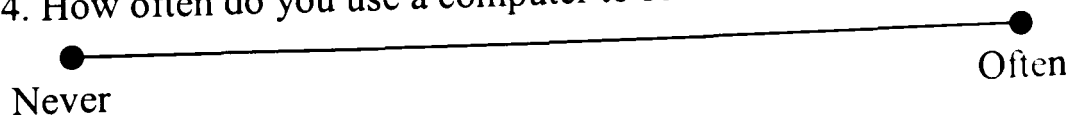


Computer section:

13. How often do you use a computer to access the Internet?



14. How often do you use a computer to send and receive email?



:: APPENDIX 11

The Principal Components Analysis study conducted to evaluate the construct validity of the Technographic questionnaire

Abstract

A Technographic questionnaire was designed to gather data about the participants' age, gender and previous telephone and computing experience. The questionnaire was divided into three sections corresponding to three components of computing experience. These sections were 'mobile phone', 'fixed line telephone' and 'computing'. The first research question concerned whether the three sections of the questionnaire were coherent internal constructs, and utilised a principal components analysis to investigate this. The results showed that the 'computing' and 'mobile phone' components were coherent constructs, but that the 'fixed line telephone' component was not. The second research question concerned whether any additional coherent components could be identified from the questionnaire. From the results of the principle components analysis, a new component consisting of a sub-group of questions from the 'mobile phone' section was identified. This new component was named 'Mobile information access (MIA)' and was composed of questions relating to mobile information access, for example, accessing voice messages. The third research question concerned whether there was any association between the new component MIA, and its 'parent' component 'mobile phone'. There was found to be a positive correlation.

1. Introduction

1.1 Research questions

The Technographic questionnaire initially required participants to provide their age and gender, followed by 16 questions relating to previous computing experience. These questions were divided into three sections corresponding to three separate areas, or components, of computing experience. However it is possible that questions from different sections, or within the same sections, correlated highly together, and could represent components, or factors, not initially recognised during the questionnaire design. Based upon this rationale, the first research question that will be addressed within this paper is:

Question 1: Is the questionnaire designed according to coherent internal constructs?

The second research question addresses the possibility that there may be a cluster of questions within the questionnaire that represent a different component of computing experience, or a sub-component of the three original components:

Question 2: Aside from the three components that the questionnaire was based upon, are there any latent components that may be coherently identified?

If a further component of the questionnaire was discovered, and it was a sub-component of one of the original three components, it would be valuable to know how it correlated with the original section. If a positive correlation existed between the original component (e.g. mobile phone) and the sub-component, this would allow the sub-component to coherently co-exist within the original component. The third research question will address this issue:

Question 3: Is there any association between the means of the scores of sub-components with the means of the scores of the original component?

2. Method

2.1 Design

The study was a within subjects design with all participants being tested under all conditions.

2.2 Participants

45 participants from the Department of Psychology at the University of Portsmouth took part in the study. The participants were recruited informally, and included undergraduates, postgraduates and staff. Participants consisted of 30 females and 15 males with ages ranging from 18 to 41 years of age.

2.3 Procedure

The Technographic questionnaire took approximately five minutes to complete, and was administered to respondents in two main ways. Firstly, fellow PhD students and some staff were asked whether they would mind completing the questionnaire. Secondly, the questionnaire was distributed to the students who had attended an undergraduate lecture on Human Factors. The students were asked whether they would mind completing the questionnaire before they departed from the lecture theatre, but there was no obligation to do this.

2.4 Data analysis

The Technographic questionnaire used a visual analogue scale, and therefore produced interval data. A principal components analysis was performed on the data from the Technographic questionnaire. This analysis was appropriate to investigate whether the questionnaire was internally consistent, and whether there was any correlation between variables to form latent components that were not recognised as part of the questionnaire design. For example, it may have been the case that high scores on question 14 concerning computer-based email usage, correlated with high scores on questions 4 and 12 concerning keypad usage vs. speech usage when using automated telephone services. This could have been the result of participants with high email usage being more comfortable using the keypad to respond to system prompts.

3. Results

3.1 Coding the data

Each question was given a variable name (e.g. q1, q2, q3, etc...), and also a more descriptive label (see Table A11.1). The continuous scale for each question was 100mm long, therefore each response was measured with a ruler, and recorded as a number ranging from 0 (often) to 100 (never).

Table A11.1. Variable labels assigned to the Technographic questionnaire items

Question number	Variable label
1	Mb Voice calls
2	Mb Voice messages
3	Mb Automated phone services
4	Mb Keypad
5	Mb Text messages
6	Mb Mobile internet services
7	Mb Other services
8	Tel Voice calls
9	Tel Voice messages
10	Tel Fax services
11	Tel Automated services
12	Tel Keypad
13	Comp Internet
14	Comp Email
15	Comp Internet services
16	Comp Other tasks

3.2 Checking the data

In order to check the data for outliers and anomalies, descriptives analyses were run on all variables. These analyses revealed that for each question 45 responses were recorded, all of which were within the range of 0 – 100, and therefore the data were treated as being correctly recorded and inputted.

3.3 Research question one

Was the questionnaire designed according to coherent internal constructs?

To investigate this question a principal components analysis (PCA) was performed to determine whether the variables grouped together into the three components underlying the questionnaire design (mobile phone, fixed line telephone, and computer). Correlation coefficient values of less than 0.4 were suppressed in order to

make interpretation of the rotated component matrix easier. A scree plot was selected as an option, and the varimax rotation method was also selected in order to maximise the relationship between the variables. The full correlation matrix between all questions can be seen in Table A11.2.

Table A11.2. Rotated component matrix from the PCA

	Component				
	1	2	3	4	5
Comp Email	0.90				
Comp Internet	0.87				
Tel Voice calls	0.78				
Tel Keypad	0.67				
Comp Other tasks	0.65				
Mb Text messages		0.81			
Mb Other services		0.79			
Mb Voice calls		0.77			
Mb Voice messages		0.58		0.43	
Comp Internet services			0.79		
Tel Fax services			0.73		
Tel Automated services			0.67		
Mb Automated phone services				0.83	
Mb Mobile internet services				0.81	
Tel Voice messages					0.87
Mb Keypad					-0.58

From the rotated component matrix it can be seen that five components with Eigenvalues above 1 emerged. The variables that formed each of the five components with Eigenvalues above 1 are listed in Table A11.3 below.

Table A11.3. The 5 components produced by the principal components analysis

Component number	Component name	Variables within each component
1	Computer	Comp Email Comp Internet Tel Voice calls Tel Keypad Comp Other tasks
2	Mobile	Mb Text messages Mb Other services Mb Voice calls Mb Voice messages
3	Services	Comp Internet services Tel Fax services Tel Automated services
4	Mobile information access (MIA)	Mb Voice messages Mb Automated phone services Mb Mobile internet services
5	Miscellaneous	Tel Voice messages Mb Keypad

After rotation, these five components accounted for 71.41% of the total variance. The amount of variance accounted for by each individual component can be seen in Table A11.4 below.

Table A11.4. Percentage variance for each component after rotation

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of variance	Cum %	Total	% of variance	Cum %	Total	% of variance	Cum%
1	4.15	25.96	25.96	4.15	25.96	25.96	3.54	22.11	22.11
2	2.73	17.06	43.02	2.73	17.06	43.02	2.34	14.64	36.74
3	1.91	11.95	54.97	1.91	11.95	54.97	2.06	12.88	49.63
4	1.47	9.16	64.13	1.47	9.16	64.13	1.90	11.88	61.51
5	1.17	7.28	71.41	1.17	7.28	71.41	1.58	9.90	71.41
6	0.88	5.48	76.89						
7	0.68	4.23	81.12						
8	0.62	3.90	85.02						
9	0.59	3.69	88.70						
10	0.46	2.88	91.58						
11	0.40	2.48	94.06						
12	0.33	2.09	96.15						
13	0.24	1.49	97.64						
14	0.17	1.04	98.68						
15	0.16	1.00	99.68						
16	0.05	0.32	100.00						

Having identified the components and their constituent parts it was then necessary to explain what the component groupings meant by naming them. This process involved analysing the variables within each component to identify whether they could be coherently linked to the same basic domain. The components that could be coherently identified and named from this data were:

Component 1: computer use. This component will be referred to by the name 'computer'. Of the four questions concerning computer usage on the questionnaire, three appeared within this component (Comp Email, Comp Internet, Comp Other tasks). Of the other two variables within this component (Tel Voice calls, Tel Keypad), it was possible to link only one to computing (Tel Keypad). 'Tel Keypad' refers to a preference for inputting data using the keypad rather than speech, which is a similar input device to the computer keyboard.

Component 2: mobile phone usage. This component will be referred to by the name 'mobile'. All of the variables within this component were from the mobile phone section of the questionnaire, and three of the four variables (Mb Text messages, Mb Other services, Mb Voice calls, Mb Voice messages) referred to the most commonly used tasks performed with a mobile phone. These tasks are making voice calls, sending text messages, and accessing voice messages.

Component 4: mobile phone information access. This component will be referred to by the name 'mobile information access (MIA)'. The three variables within this component are all concerned with accessing information from a mobile phone. 'Mb Voice messages' refers to accessing voice messages. 'Mb Automated phone services' refers to accessing information using automated telephone services. 'Mb Mobile internet services' refers to accessing information by wirelessly connecting to the Internet.

From this analysis, only two of the new components (computer, mobile) approximate to the original three components. The questionnaire therefore demonstrates internal consistency for two of the original components, but not for 'fixed line telephone'.

The components that could not be coherently identified and named from this data were components three and five, which will be referred to as components 'services' and 'miscellaneous' respectively. These components consisted of variables that were not linked to the same base domain.

3.4 Research question two

Aside from the three components that the questionnaire was based upon, are there any latent components that may be coherently identified?

Of the three new components, component MIA consists of a sub group of variables from within the mobile phone component. The MIA component was not recognised as a separate component when designing the questionnaire, and it is possible that it could represent an independent aspect of previous computing experience.

3.5 Research question three

Is there any association between the means of the scores of sub-components with the means of the scores of the original component?

The MIA component emerged as a sub-component of the mobile phone component, consisting of three variables from the mobile phone section of the original questionnaire. It was therefore possible that an association existed between the MIA component and the mobile phone section. In order to investigate this, and to investigate any associations that may exist between the other two sections of the questionnaire, bivariate correlations were performed on the means of the three original components of the questionnaire, and on the MIA component. Table A11.5 shows the means and standard deviations.

Table A11.5. Descriptives data for the PCA components

Component	Mean	Standard deviation
1. Mean of mobile phone section	65.64	14.90
2. Mean of fixed line telephone section	63.42	17.22
3. Mean of computer section	27.56	17.10
4. Mean of MIA component	84.29	13.20

There was a significant correlation between the MIA component and the mobile phone section ($r=0.66$; $n=45$; $p<0.01$). Therefore participants who scored highly on the mobile phone section are also likely to score highly for the MIA component. This result demonstrates that the mobile phone component may be separated into two separate components, each of which represents an aspect of mobile phone use, and the scores of which will be correlated.

4. Discussion

The results will now be discussed in relation to each of the initial research questions.

4.1 Research question one

Was the questionnaire designed according to coherent internal constructs?

Aside from gender and age, the questionnaire was primarily designed to record participants' previous experience with computing devices. The three areas, or

components, of computing experience chosen were mobile phone, fixed line telephone, and computing. These three components appear to be clearly separated, but with the convergence of many functions into one device (e.g. mobile phones allow text messaging and internet access in addition to call capability) there may be overlapping areas of functionality, or more specific areas of functionality within one of the main three components. It is for this reason that the original three components were analysed using principal components analysis to determine whether they emerged as coherent internal constructs. In other words, whether they were valid as separate components of computing experience. The results showed that the 'computing' and 'mobile phone' components were valid components, but that the 'fixed line telephone' component did not show any coherency as a separate component.

On inspection of Table A11.2 it can be seen that the fixed line telephone service variables were separated into various different components. Future work could therefore involve redesigning this section as two sub-sections. The first section could consist of basic, commonly used tasks like calling and retrieving voice messages. The second section could consist of more advanced, less commonly used tasks such as accessing automated services, and using fax.

4.2 Research question two

Aside from the 3 components that the questionnaire was based upon, are there any latent components that may be coherently identified?

The principal components analysis produced five components, three of which were coherently named (computer, mobile, MIA). Of the three components two were consistent with two of the original questionnaire sections (computer, mobile). The final component was named MIA and consisted of a sub-group of variables from the mobile phone component.

4.3 Research question three

Is there any association between the means of the scores of sub-components with the means of the scores of the original component?

This question aims to discover whether the sub-component (MIA) of the mobile phone section produced similar scores to the section as a whole, or whether it produced very different scores. A strong association existed between the main component and its sub-component (MIA), which means that if the section were divided into two separate areas, the scores for each would correlate, even though they were targeting different aspects of mobile phone usage. The mobile phone section of the questionnaire therefore measures two separate aspects of mobile phone experience, one of which is concerned with information access using telephone services (MIA), and the other of which is concerned with the communicative aspects of the device (text messages and voice calls). However, for the purposes of the work reported in this thesis the mobile phone section will be used to measure all aspects of previous mobile phone experience.

5. Conclusions

Based on the analysis of the questionnaire responses, the technographic questionnaire can be used to reliably measure previous levels of mobile phone, and computing experience. For future directions for development of the questionnaire include increasing the number of questionnaires administered, and analysed using the principal components analysis technique, in order to achieve more stable and reliable components. It may also be necessary to include goodness of fit statistics, such as KMO and Bartlett's test of sphericity values, in order to determine if the five-factor model is reliable.

:: APPENDIX 12

The task sheet for preliminary study two

The practice task, and 3 experimental tasks for the standard service:

- **Practice task:** Find the names of 2 Pubs that close at 1am.
- **Task 1:** Find the names of 2 Wine Bars that close at 11pm.
- **Task 2:** Find the names of 2 budget Chinese restaurants, and then exit the service.
- **Task 3:** Find the names of 2 Café Bars that close late, and then find the names of 2 nightclubs playing Techno music, and then exit the service.

The 3 experimental tasks for the metaphor-based service:

- **Task 1:** Find the names of 2 top-end Italian restaurants.
- **Task 2:** Find the names of 2 Jazz cafes that close at 8pm, and then exit the service.
- **Task 3:** Find the names of 2 budget Indian restaurants, and then find the names of 2 Hip Hop Nightclubs, and then exit the service.

:: APPENDIX 13

The practice task sheet, and task sheet for experiment one

Session 1:

- **Practice task:** Find the names of 2 nightclubs playing Hip Hop music, and then exit the service.
- **Task 1:** Find the names of 2 wine bars that close at 11pm and then exit the service.
- **Task 2:** Find the names of 2 science fiction films showing late night, and then find the names of 2 thriller films showing late night, and then exit the service.
- **Task 3:** Find the names of 2 mid-range Italian restaurants, and then find the names of 2 comedy films showing in the evenings, and then exit the service.

Session 2:

- **Task 1:** Find the names of 2 budget Chinese restaurants and then exit the service.
- **Task 2:** Find the names of 2 cyber cafes that close at 8pm, and then find the names of 2 café bars that close late, and then exit the service.
- **Task 3:** Find the names of 2 pop art exhibitions that are open weekdays, and then find the names of 2 jazz cafés that close at 5pm, and then exit the service.

Session 3:

- **Task 1:** Find the names of 2 expressionist art exhibitions that are open at weekends and then exit the service.
- **Task 2:** Find the names of 2 pubs that close at 1am, and then find the names of 2 nightclubs playing techno music, and then exit the service.
- **Task 3:** Find the names of 2 thriller films showing in the evenings, and then find the names of 2 pop concerts, and then exit the service.

Session 4:

- **Task 1:** Find the names of 2 artists playing live blues music and then exit the service.
- **Task 2:** Find the names of 2 top-end Chinese restaurants, and then find the names of 2 top end Indian restaurants, and then exit the service.
- **Task 3:** Find the names of 2 abstract art exhibitions that are open in the evenings, and then find the names of 2 pre-club bars that close at 11pm, and then exit the service.

:: APPENDIX 14

The post-task informal interview questions for experiment one

1. What features did you like about the service?
2. What features didn't you like about the service?
3. What did you think about the different characters in the service?
4. Generally, did you find the service easy or hard to use?
5. Is there anything that you would add to the service to make it easier to use?
6. If you could change one aspect of the service what would it be?
7. What did you think about the tasks you had to do?
8. Where were you when you took part in this experiment?

:: APPENDIX 15

The post-task informal interview questions for experiment two

1. What features did you like about the service?
2. What features didn't you like about the service?
3. If you could change one aspect of the service what would it be?
4. Is there anything that you would add to the service to make it easier to use?
5. Were you able to visualise the structure of the service?
6. How well did you remember the structure of the service from week to week?
7. How did you feel using the service in a private place?
8. How did you feel using the service in a public place?
9. What were the main differences between using the service in a private and a public place?
10. How does the office filing system service compare to the service that you used in the first week?
11. What did you think about the voice?

:: APPENDIX 16

The mobile phone attitude questionnaire

Please read each statement carefully, and then using the 7-point rating scale below, tick the box that best represents your agreement or disagreement with the statement.

Strongly agree	Agree	Slightly agree	Neutral	Slightly disagree	Disagree	Strongly disagree
1	2	3	4	5	6	7

No	Statement	1	2	3	4	5	6	7
1	I like the way mobile phones make people constantly available to others							
2	I feel comfortable being around other people when I use a mobile phone							
3	I feel awkward answering a mobile phone call when I am in a busy public place							
4	I feel it is rude when people use their mobile phones in busy public places							
5	I feel uncomfortable when people are talking on mobile phones near me in public places							
6	I do not find it irritating when I hear mobile phones ring in public places							
7	When I use a mobile phone the only time I feel I have enough privacy is when no one else is around							
8	I do not feel excluded when friends use their mobile phones when I am with them in a public place							
9	I feel comfortable using my mobile phone in a busy public place							
10	I do not like the feeling of being watched when I use my mobile phone in public places							
11	I always carry my mobile phone with me wherever I go							
12	I feel vulnerable when using a mobile phone in a							

	busy public place								
13	I prefer texting (SMS) people rather than phoning them when I am in a busy public place								
14	I feel too many people make unimportant mobile phone calls in public places								
15	I feel comfortable when people listen to my conversation when I am using a mobile phone in a public place								
16	I believe mobile phones should only be used for emergencies								
17	When using a mobile phone other people's behaviour nearby makes me nervous								
18	I do not feel embarrassed when my mobile phone rings in a busy public place								
19	I believe there is generally enough space around me when I use a mobile phone								
20	I do not feel I have to hurry when using a mobile phone in a busy public place								

:: APPENDIX 17

The task sheet for experiment two

Control Session:

- **Practice task:** Find the names of 2 nightclubs playing Hip Hop music, and then exit the service.
- **Task:** Find the names of 2 wine bars that close at 11pm and then exit the service.

Session 1:

- **Task:** Find the names of 2 science fiction films showing late night, and then find the names of 2 thriller films showing late night, and then exit the service.

Session 2:

- **Task:** Find the names of 2 top-end Chinese restaurants, and then find the names of 2 top end Indian restaurants, and then exit the service.

Session 3:

- **Task:** Find the names of 2 pubs that close at 1am, and then find the names of 2 nightclubs playing techno music, and then exit the service.

Session 4:

- **Task:** Find the names of 2 comedy films showing in the evenings, and then the names of 2 abstract art exhibitions that are open at the weekends

Session 5:

- **Task:** Find the names of 2 mid-range Italian restaurants, and then find the names of 2 pop concerts, and then exit the service.

Session 6:

- **Task:** Find the names of 2 pop art exhibitions that are open weekdays, and then find the names of 2 jazz cafés that close at 5pm, and then exit the service.

:: APPENDIX 18

The post-task informal interview questions for experiment three

1. What features did you like about the service?
2. What features didn't you like about the service?
3. Did you have any problems using the service?
4. Is there anything that you would add to the service to make it easier to use?
5. Were you able to visualise the structure of the service?
6. What aspects of the service did you visualise?
7. What strategy did you use for navigating through the service?
8. Did you always know where you were within the service?
9. What features of the service helped you to know where you were?
10. How well did you remember the structure of the service from task to task?
11. How did you feel using the service for the first time?
12. At what point did you start to feel confident using the service?
13. How does this service compare to the standard menu style service?
14. How did you feel using the service in a public place?

:: APPENDIX 19

The multiple choice memory questionnaire

Please select one answer for each of the following questions by ticking the appropriate box.

1. How many levels of menu options were there in the service?

2 3 4 5 6

The following questions 2-7 refer to the options available for navigating forwards through the service, and EXCLUDE the options: Repeat, Back, Return, Exit, and Help. The number of questions you answer will depend on your answer to question 1.

2. How many menu options were available at the first level of the service?

1 2 3 4 5

3. How many menu options were available at the second level of the service?

1 2 3 4 5

4. How many menu options were available at the third level of the service?

1 2 3 4 5

5. How many menu options were available at the fourth level of the service?

1 2 3 4 5

6. How many menu options were available at the fifth level of the service?

- 1 2 3 4 5

7. How many menu options were available at the six level of the service?

- 1 2 3 4 5

:: APPENDIX 20

The task sheet for experiment three

- **Practice task:** Find the names of 2 nightclubs playing Hip Hop music, and then exit the service.
- **Task 1:** Find the names of 2 wine bars that close at 11pm, and then exit the service.
- **Task 2:** Find the names of 2 budget Chinese restaurants, and then find the names of 2 budget Indian restaurants, and then exit the service.
- **Task 3:** Find the names of 2 cyber cafes that close late, and then find the names of 2 pubs that close at 1am, and then exit the service.

:: APPENDIX 21

Absolute transaction times

Tables A21.1 to A21.4 show the mean times taken to perform all tasks within an experimental trial by participants using the standard service, and by participants using metaphor-based services.

Table A21.1. Absolute times for preliminary study two

Service type	Mean time (s)	SD
Standard	359.05	104.12
Metaphor-based	466.35	130.22

Table A21.2. Absolute times for experiment one

Service type	Mean time (s)	SD
Standard	290.82	40.62
Metaphor-based	439.50	84.38

Table A21.3. Absolute times for experiment two

Service type	Mean time (s)	SD
Standard	143.26	18.20
Metaphor-based	174.82	15.88

Table A21.4. Absolute times for experiment three

Service type	Mean time (s)	SD
Standard	137.51	30.71
Metaphor-based	147.55	31.40