SpaceBook: Designing and Evaluating a Spoken Dialogue Based System for Urban Exploration

William Mackaness[1], Phil Bartie[1], Tiphaine Dalmas[2],
Srini Janarthanam[3], Oliver Lemon[3], Xingkun Liu[3] and Bonnie Webber[3]

[1]Sch of Geosciences, [2]Sch Informatics, University of Edinburgh, EH8 9XP
[3]School of Mathematics and Computer Science, Heriot-Watt University Edinburgh
william.mackaness@ed.ac.uk, philbartie@gmail.com, tiphaine.dalmas@aethys.com, Srini Janarthanam srinivasancj@gmail.com, O.Lemon@hw.ac.uk, xingkun.liu@gmail.com, bonnie@inf.ed.ac.uk

KEYWORDS: location based services, dialogue based interaction, urban models, cognitive models, visibility analysis

1. The Essence of SpaceBook: 'Hands Free, Eyes Free'

When tourists travel to new cities, they desire an authentic experience, wishing to explore and learn about the environments in which they find themselves. Clearly this is not an easy task since we can readily spot tourists looking quizzically at a crumpled map, whilst browsing *Lonely Planet* guide books, and trying not to look conspicuous whilst taking numerous snapshots. Location-aware smartphones are now an important supplement to that toolkit (Ishikawa et al. 2008), providing a range of services (on-screen maps, answers to 'where is my nearest', and access to the web and thence its many services) (Krug et al 2003; Cheverst et al. 2000; Jost et al 2005; Zipf and Jost 2005). We argue that the process of smartphones interaction is distracting – requiring complex hand-and-eye coordination over hard-to-read screens, as well as interpretive map skills, that collectively distract the tourist from the pleasure of the city. Instead we advocate a concealed technology, in which interaction is solely speech-based, leaving the tourist 'eyes and hands free', able to ask about anything around them or that comes into their field of view as they explore the city. This is the vision of SpaceBook. This paper builds on work in 'interaction via the non visual' (Jeong and Gluck, 2003; Jacob et al. 2010; Bartie and Mackaness 2006). It describes SpaceBook's essential building blocks and reports on the evaluation of a prototype, in comparison to existing technologies used for navigation and search.

2. The SpaceBook System

Hands-free, eyes-free responsive support for urban exploration demands a dialogue-based system that is capable of both responding to the tourist's requests and calling their attention to interesting visible features of their surroundings. It requires that we handle vague and multi scale representations of space (Montello et al. 2003). This is the task of the interaction manager (IM) (Janarthanam et al. 2012). Processing the tourist's utterances requires automatic speech recognition (ASR) to convert the audio into text, and text to speech (TTS) software so that constructed utterances can be delivered back as audio. In order to give context to the spoken word (Lemon and Gruenstein 2004), we need a model of the city (CM) that is being explored and information of the tourist's location. The city model also needs to model 3D space, so that we can work out what is in the field of view at any given instant. A pedestrian tracker module (PT) is used to determine their location at any given point in time,

and record where they have been. If we are to answer the tourist's questions, we need a question answering component (QA). Figure 1 summarises connections between these components.

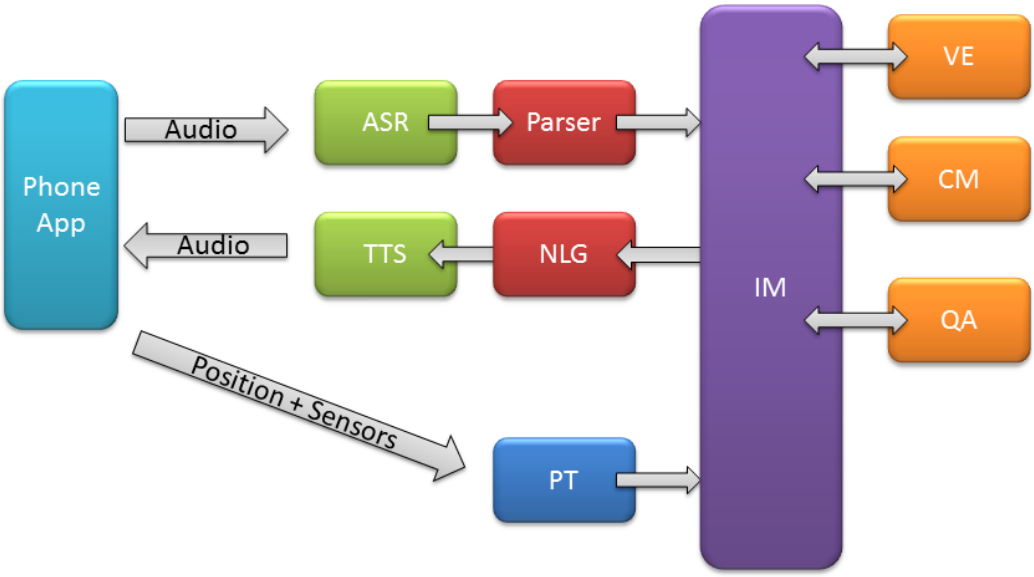

Figure 1. SpaceBook System

Phone App: The current version of the system runs on a Samsung Galaxy S3 / Note Android phone which has access to both GPS and GLONASS GNSS. It sends this position data and other sensor data to the pedestrian tracker component. The PhoneApp also transmits audio data to the ASR component, and receives audio data from the TTS component.

ASR: The Automatic Speech Recognition is handled by Nuance 9.0 and a Freeswitch interface, using a grammar-based Language Model. The grammar and vocabulary includes word lists for the street, entities, and entity-types within the test area, as well as prominent people and places that may be the subject of Question-Answering inputs (e.g. Mary Queen of Scots, Harry Potter, Edinburgh Castle). The grammar consists of approximately 80 rules, covering user navigation goal inputs, Question-Answering inputs, visibility statements, and general dialogue-management inputs.

Parser: The parser module translates the user utterances from the ASR module into a form that can be understood by the Interaction Manager (IM).

PT: The Pedestrian Tracker module combines the sensor and positioning information with spatial data from the city model (CM) to calculate the most likely position of the pedestrian, improving upon the raw GNSS output. It also estimates the mode of transport through analysis of the sensor data.

IM: The Interaction Manager is the central module that manages the user's interaction with the system (Janarthanam et al. 2012). It receives the user's location coordinates from the Pedestrian Tracker and the user's utterances from the ASR module. It then sends recognised

user utterances (as strings) to the parser, which translates each one into a semantic representation (called a *dialogue act*) in SpaceBook's meaning representation language. Based on the dialogue act and the user coordinates, the IM decides the next course of action.

Overall, the IM manages three tasks:

1. Navigating the user:
2. Pushing information about Points of Interest:
3. Exploration using Question Answering (QA): Open questions from the user such as "Who is David Hume?", "When was Mary Queen of Scots born?".

QA: The Question Answering component finds information relevant to open requests from the user. These open requests might be for the descriptions of things, biographical information and additional information about anything they have heard (*pull behaviour*) and to IM queries (*push behaviour*) about the user's surroundings.

CM: The City Model is a spatial database holding information on networks, points of interest, and polygon boundary definitions for features of interest and is accessed using spatial SQL.

VE: The visibility engine computes which parts of the city are in view to the pedestrian at any given time. It was built on a DSM and a DEM created from LiDAR data. A range of visibility metrics are output, including the field of view occupied, and the amount of facade area visible (Bartie et al 2010).

NLG: The Natural Language Generation component takes a content planning input from the IM and realizes it in English.

TTS: CereProc (http://www.cereproc.com/) was used as the Text-to-Speech Engine. It was used with a Scottish female voice called 'Heather'.

3. Evaluation of SpaceBook

A number of evaluative methodologies exist for dialogue based systems (Lemon et al 2006). Evaluation of the initial SpaceBook prototype was undertaken in Edinburgh in August 2012 and compared against Samsung S-voice and Google Voice Search for navigation and exploration questions. S-voice is a central voice application from which the user can launch navigation or search, e.g. "directions to Edinburgh castle" will start Google navigation, "search David Hume" will launch a Google search for David Hume. Users were briefed on these systems prior to the experiments. Users evaluated both systems in a walk-about through the city that involved them carrying out specific tasks. The walk-about was divided into two legs. On the first leg, they used one system and then swapped to the second system on the second leg.

3.1 User Tasks

Users were asked to be guided to restaurants and a sports centre, to ask historical questions and for information about historical landmarks that they passed (Figure 2). Leg two was similar in ambition, but with a different set of questions.

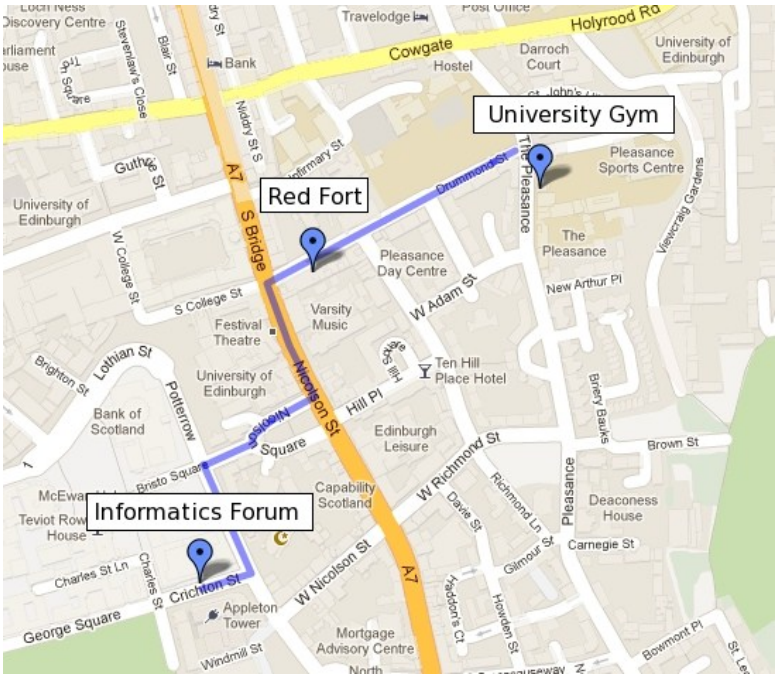

Figure 2. Leg 1 of User Task (shown on Google Maps background)

3.2 Analysis

The following information was gathered as a basis for identifying future improvements:

a) voice recording of user and system utterances in order to measure accuracy of ASR, and asking subjects to comment on their actions when using Google navigation;

b) Sensor and Position Data streaming GNSS and accelerometer data to explore positional latency and precision;

c) Subject and System Utterances;

All utterances were time-stamped, in order to assess the speed of the system overall. The range of utterances included:

Greetings: e.g., “My name is spacebook. I am waiting for your location information. Please wait.”, or “I am ready. How can I help you?”

Responses to navigational questions: e.g., “the street is about 4 kilometres away.”, “I will give you further instructions automatically when you are near a junction.”

Requests for clarification: e.g., “Did you say, who is James Watt?”

Navigational information: e.g., “You are on Saint Mary's Street.”, “You will walk past Post Office on your left.”

***Pushed* exploratory information**: e.g., “Located in Hyndford's Close off Edinburgh's High Street, opposite the Brass Rubbing Centre and John Knox's House is the Museum of Childhood, which has been described as ""the noisiest museum in the world"". Its collection includes historic games and toys, including dolls and teddy bears, from many countries.”

Responses to open questions: e.g., "I searched for John Knox. Gazetteer for Scotland says Churchman. Born in Haddington, East Lothian, Knox was known as the father of the Protestant Reformation in Scotland, whereby the Roman Catholic Church was replaced by a disestablished, democratic, Presbyterian Church of Scotland, founded on Calvinist principles."

Problem indicators: e.g., "Looks like I cannot track you on GPS. Could you hang up and call again?"

Just prior to and just after the experiments, the user completed a questionnaire. The user was followed to record their responses to information and to assess the efficacy of the route following instructions, the ambient conditions, and what features in the landscape they looked at. Results indicated a preference for Google's speech and for Google's navigation when a choice had to be made between the two systems. SpaceBook's exploration aspect was preferred and subjects also liked the novelty of the SpaceBook interface, rating the interaction more interesting than Google.

4. Future Developments of SpaceBook

SpaceBook is far from complete and experiments identified the need for improvements in speech recognition, text to speech production, and the handling of uncertainty. A balance needs to be found between the push and pull of information, and the need to prioritise information at critical decision points along the journey.

5. Acknowledgements

The research leading to these results has received funding from the EC's 7th Framework Programme (FP7/2011-2014) under grant agreement no. 270019 (SpaceBook project).

References

Bartie P.J. and Mackaness W.A. Development of a speech-based augmented reality system to support exploration of cityscape. Transactions in GIS, 10(1):63–86, 2006.

Bartie P, Reitsma F, Kingham S, Mills S Advancing visibility modelling algorithms for urban environments. Computers Environment and Urban Systems 34: 518-531, 2010

Cheverst K., Nigel Davies, Keith Mitchell, and Paul Smith. Providing tailored (context-aware) information to city visitors. In Proceedings of First International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, pages 73–85, 2000.

Ishikawa, T., Fujiwara, H., Imai, O. and Okabe, A. (2008). Wayfinding with a gps-based mobile navigation system: maps and direct experience, *Journal of Environmental Psychology* 28(1): 74–82.

Jacob, R., Mooney, P., Corcoran, P. and Winstanley (2010). Haptic-gis: Exploring the possibilities, *SIGSPATIAL Special* 2: 36–39

Jeong, W. and Gluck, M. (2003). Multimodal geographic information systems: adding haptic and auditory display, *J. Am. Soc. Inf. Sci. Technol.* 54: 229–242.

J¨ost M., Jochen H¨aussler, Matthias Merdes, and Rainer Malaka. Multimodal interaction for pedestrians: an evaluation study. In IUI '05: Proceedings of the 10th international conference on Intelligent user interfaces, pages 59–66, 2005.

Janarthanam, S, Lemon, O, Xingkun Liu, X., Bartie, P, Mackaness, W.A., Dalmas, T and Goetze J., "Integrating Location, Visibility, and Question-Answering in a Spoken Dialogue System for Pedestrian City Exploration", Proceedings of SIGDIAL 2012

Krug, K., David Mountain, and Debbrah Phan. Webpark: Location-based services for mobile users in protected areas. GeoInformatics, pages 26–29, March 2003.

Lemon, O, Kallirroi Georgila, and James Henderson. Evaluating Effectiveness and Portability of Reinforcement Learned Dialogue Strategies with real users: the TALK TownInfo Evaluation. In IEEE/ACL Spoken Language Technology, 2006.

Lemon, O. and Alexander Gruenstein. Multithreaded context for robust conversational interfaces: context-sensitive speech recognition and interpretation of corrective fragments. ACM Transactions on Computer-Human Interaction (ACM TOCHI), 11(3):241–267, 2004.

Montello, D. R., Goodchild, M. E, Gottsegen, J., & Fohl, P. (2003). Where's Downtown?: Behavioural methods for determining referents of vague spatial queries. Spatial Cognition and Computation, 3(2&3), 185-204.

Rayner M. Lewin I. Becket R. Carter D. Boye, J. and M. Wiren. Language processing strategies and mixed-initiative dialogues. Electronic Transactions of Artificial Intelligence, 3(D):73–88, 1999.

Zipf A. and M. J¨ost. Implementing adaptive mobile GI services based on ontologies - examples for pedestrian navigation support. Computers, Environment and Urban Systems - An International Journal. Special Issue on LBS and UbiGIS., 2005

Biography

William Mackaness is a lecturer in the School of GeoSciences at The University of Edinburgh. Phil Bartie completed his PhD in Camterbury (NZ) and is a Research Associate in GeoSciences working on the SpaceBook project, as is Tiphaine Dalmas (School of Informatics, University of Edinburgh), Srini Janarthanam, and Xingkun Liu (both working as Research Associates in the Interaction Lab at Heriot-Watt University). Oliver Lemon is a Professor at the School of Mathematics and Computer Science, Heriot-Watt and -Bonnie Webber is a Professor in the School of Informatics. Their collective skills cover location based services, 3D visibility analysis, Natural Language processing and semantics, dialogue management, natural language generation, user modelling, knowledge representation and machine learning, open-domain question answering, technology enhanced learning and goal inferencing from spoken utterances.