A Test-Bed for Text-to-Speech-Based Pedestrian Navigation Systems

Michael Minock[1], Johan Mollevik[2], Mattias Åsander[2], and Marcus Karlsson[2]

[1] School of Computer Science and Communication (CSC),
Royal Institute of Technology KTH, Stockholm, Sweden
[2] Department of Computing Science, Umeå University, Umeå, Sweden

Abstract. This paper presents an Android system to support eyes-free, hands-free navigation through a city. The system operates in two distinct modes: *manual* and *automatic*. In manual, a human operator sends text messages which are realized via TTS into the subject's earpiece. The operator sees the subject's GPS position on a map, hears the subject's speech, and sees a 1 fps movie taken from the subject's phone, worn as a necklace. In automatic mode, a programmed controller attempts to achieve the same guidance task as the human operator.

We have fully built our manual system and have verified that it can be used to successfully guide pedestrians through a city. All activities are logged in the system into a single, large database state. We are building a series of automatic controllers which require us to confront a set of research challenges, some of which we briefly discuss in this paper. We plan to demonstrate our work live at NLDB.

1 Introduction

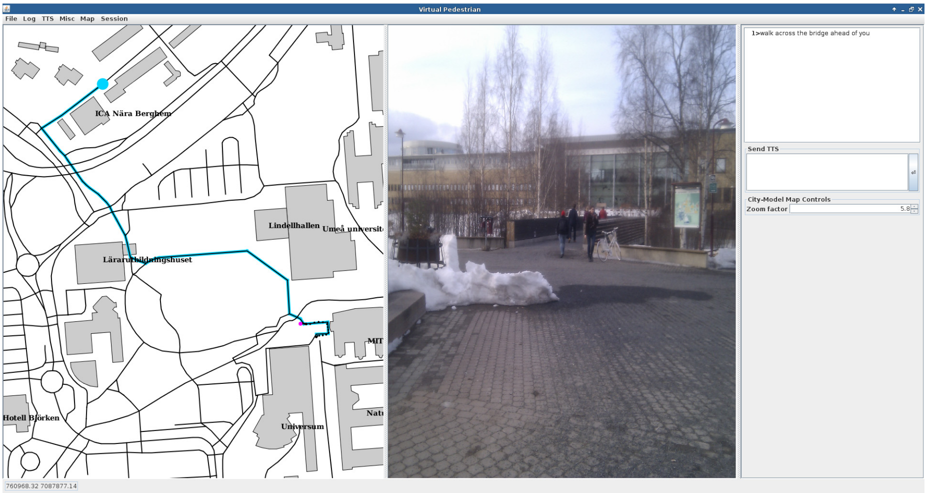
The automated generation of route directions has been the subject of many recent academic studies (See for example the references in [1], or the very recent works [2,3]) and commercial projects (e.g. products by Garmin, TomTom, Google, Apple, etc.). The pedestrian case (as opposed to the automobile case) is particularly challenging because the location of the pedestrian is not just restricted to the road network and the pedestrian is able to quickly face different directions. In addition, the scale of the pedestrian's world is much finer, thus requiring more detailed data. Finally the task is complicated by the fact that the pedestrian, for safety, should endeavor to keep their eyes and hands free – there is no room for a fixed dashboard screen to assist in presenting route directions. We take this last constraint at full force – in our prototype there is no map display; the only mode of presentation is text-to-speech instruction heard incrementally through the pedestrian's earpiece.

We present a system to support eyes-free, hands-free navigation through a city[1]. Our system operates in two distinct modes: *manual* and *automatic*. In manual

[1] The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 270019 (SpaceBook project `www.spacebook-project.eu`) as well as a grant through the Kempe foundation (`www.kempe.com`).

E. Métais et al. (Eds.): NLDB 2013, LNCS 7934, pp. 396–399, 2013.

mode, an operator guides a subject via text-to-speech commands to destinations in the city. The operator, working at a stationary desktop, receives a stream of GPS, speech and camera image data from the subject which is displayed in real time to the operator (see figure 1). In turn the operator types quick text messages to guide the subject to their destination. The subject hears the operator's instructions via the text-to-speech engine on their Android. In automatic mode the human operator is missing, replaced by a programmed controller.



1>walk across the bridge ahead of you

Fig. 1. Operator's interface in manual mode guiding a visitor to ICA Berghem, Umeå

The technical specification and design of our system, with an initial reactive controller, is described in a technical report [1]. That report gives a snap-shot of our system as of October 2012. In the ensuing months we have worked to optimize, re-factor and stabilize the system in preparation for its open source release – working name JANUS (Interested readers are encouraged to contact us if they wish to receive a beta-release). We have also further developed the infrastructure to integrate FreeSWITCH for speech and some extra mechanism to handle image streams. Finally we have added a facility that logs phone pictures to PostgreSQL BLOBs, the TTS messages to PostgreSQL text fields, and the audio-streams to files on the file system. Aside for server-side PL/pgSQL functions, the system is written exclusively in Java and it uses ZeroC ICE for internal communication. Detailed install instructions exists for Debian "wheezy".

2 Field Tests

We have carried out field tests since late Summer 2012. The very first tests were over an area that covered the Umeå University campus extending North to Mariahem (An area of roughly 4 square kilometers, 1788 branching points, 3027 path segments, 1827 buildings). For a period of several weeks, the first author tested the system 3-4 times per week while walking or riding his bicycle to work and back. The system was also tested numerous times walking around the Umeå University campus. A small patch of the campus immediately adjacent to the MIT-Huset was authored with explicit phrases, overriding the automatically generated phrases of a primitive NLG component (see the example in [1]). These initial tests were dedicated to validating capabilities and confirming bug fixes and getting a feel for what is and is not important in this domain. For example problems like the quantity and timing of utterances (too much or too little speech, utterances issued too late or too early) and oscillations in the calculation of facing direction led to a frustrating user experience. Much effort was directed toward fixing parameters in the underlying system, adding further communication rules and state variables, etc.

In addition to these tests, in November 2012 we conducted an initial test of our manual interface in Edinburgh (our database covered an area of roughly 5 square kilometers, 4754 branching points, 9082 path segments, 3020 buildings) – walking the exact path used in the Edinburgh evaluations of the initial SpaceBook prototype developed by SpaceBook partners Heriot-Watt and Edinburgh University [2]. With the PhoneApp running in Edinburgh and all back-end components running in Umeå, the latencies introduced by the distance did not render the system inoperable. Note that we did not test the picture capability at that time, as it had not yet been implemented.

Due to the long Winter, we have conducted only a few outdoor tests with the system from November 2012 to April 2013. What experiments we have run, have been in an area surrounding KTH in Stockholm (An area slightly over 2 square kilometers, 1689 branching points, 3097 path segments, 542 buildings), the center of Åkersberga, and continued tests on the Umeå University campus. With the warming of the weather we look forward to a series of field tests and evaluations over the Spring and Summer of 2013.

3 System Performance

Our optimization efforts have been mostly directed at minimizing latencies and improving the performance of map rendering in our virtual pedestrian/tracking tool. There are three latencies to consider from the PhoneApp to the controller (GPS report, speech packet, image) and one latency to consider from the controller to the PhoneApp (text message transmission). We are still working on reliable methods to measure these latencies and, more importantly, their variability. In local installations (e.g. back-end components and PhoneApp running in Umeå) the system latencies are either sub-second or up to 1-2 seconds – a perfectly adequate level of performance. Running remotely (e.g. back-end components running in Umeå and PhoneApp in Edinburgh) appears to simply add a fixed constant to all four latencies.

All the map data is based on XML exports of OpenStreetMap data converted to SQL using the tool osm2sb (see [1]). We have limited our attention to what may be downloaded as XML exports via OpenStreetMap's web-site. This has covered large enough portions of the city for our purposes. That said, we strongly believe that inefficient access to larger maps is not a significant risk.

4 Some Future Challenges

Evaluations: We have a very natural metric of evaluation: *what is a controller's effectiveness in actually guiding pedestrians from a given initial position to a given destination position?* To minimize expense, we will first employ what we term *auto evaluation*. In auto evaluation one generates random tours, unknown to the subject, over a large number of possible destinations. Because destinations are hidden, even if one of the authors serves as a subject, we will gain insight into the relative effectiveness of various controller strategies. Only after performing this cheaper form of evaluation shall we carry out a larger classical evaluations with testable hypotheses, large cohorts of random subjects, control groups, etc.

Scheduling of Utterances in Synchronization with User Position: Early in our testing we found that scheduling of utterances in synchronization with user position is a critical capability that is not easily finessed in a reactive controller. Thus we have started work on the challenging problem of predicting user position and scheduling utterances accordingly. This is briefly discussed in [4] and will be presented in greater detail in a future conference paper.

Reuse of Operator Utterances in Automatic Controllers: Our current controllers fetch pre-compiled utterances populated via primitive NLG routines run off-line. While we will explore techniques to integrate run-time NLG systems, we are interested in techniques to re-use utterances expressed by human operators (in manual mode) within our automatic controllers. We seek large collections of human authored utterance variations, where, given a large number of user trials, we might learn a policy to select when and where to issue utterances to maximize expected utility over our metric of evaluation.

References

1. Minock, M., Mollevik, J., Åsander, M.: Towards an active database platform for guiding urban pedestrians. Technical Report UMINF-12.18, Umeå University (2012), `https://www8.cs.umu.se/research/uminf/index.cgi?year=2012&number=18`
2. Janarthanam, S., Lemon, O., Liu, X., Bartie, P., Mackaness, W., Dalmas, T., Goetze, J.: A spoken dialogue interface for pedestrian city exploration: integrating navigation, visibility, and question-answering. In: Proc. of SemDial 2012, Paris, France (September 2012)
3. Boye, J., Fredriksson, M., Götze, J., Gustafson, J., Königsmann, J.: Walk this way: Spatial grounding for city exploration. In: Proc. 4th International Workshop on Spoken Dialogue Systems, IWSDS 2012, Paris, France (November 2012)
4. Minock, M., Mollevik, J.: Prediction and scheduling in navigation systems. In: Proceedings of the Geographic Human-Computer Interaction (GeoHCI) Workshop at CHI (April 2013)