

ON MULTIMODAL ROUTE NAVIGATION IN PDAS

Topi Hurtig, Kristiina Jokinen
University of Helsinki, Finland

Abstract

One of the biggest obstacles in building versatile natural human-computer interaction systems is that the recognition of natural speech is still not sufficiently robust, especially in mobile situations where it's almost impossible to cancel out all irrelevant auditory information. In multimodal systems the possibility to disambiguate between several input and output modalities can substantially increase the intelligibility of dialogues and the robustness of interaction. The combination of natural speech and tactile gestures as input mediums, especially in map-based systems, have shown prominent results, although mature commercial applications are still to be developed. In this paper we present the MUMS Multimodal Route Navigation System, intended for public transportation commuting. The system allows users to present route queries with any preferred combination of speech and pen input, and the system provides navigational information via speech and graphical map representations. The focus of this document is on the system's natural interaction model, which is designed keeping in mind the current limitations in natural speech recognition.

Keywords: human-computer interaction, multimodal dialogue systems, natural language, modality fusion, user interfaces, mobile systems, route navigation

1. Introduction

Multimodal interactive systems have gained ground in recent years, and they do seem to provide a user-friendly alternative to several application fields. However, due especially to the lack of robustness in speech recognition, there is still a long way to go before any of these kind of applications pass as their human counterparts. The naturalness feature can be attached, not only to human-human communication, but also to applications that take advantage of users' natural ways of giving and receiving information. Natural interaction does not also only include verbal communication: much of the information content in human-human situations is conveyed by non-verbal signs, gestures, facial expressions, etc. Thus, in order to develop next generation human-computer interfaces, it is necessary to work on technologies that allow multimodal natural interaction: it is important to investigate coordination of natural input modes (speech, pen, touch, eye movement, etc.) as well as multimodal system output (speech, sound, graphics, etc.), ultimately aiming at intelligent interfaces that are aware of the context and user needs, and can utilize appropriate modalities to provide information tailored to a wide variety of users. Natural interaction could be considered an approach by which various users in different situations could exploit the strategies they have learnt in human-human communication.

2. Related research

Speech and tactile input are known to be very closely coupled, and their combined use has been extensively studied. For example in studies conducted with the QuickSet system (Oviatt et al. 2000; Oviatt 2001) it has been found that multimodal input can indeed help in disambiguating input signals, which improves the system's robustness and performance stability. Other advantages of multimodal interfaces include the ability to choose an input approach best suited for each person and each situation. Different modalities offer different benefits, and also the freedom of choice (Gibbon et al. 2000). Jokinen and Raïke (2002) also point out that multimodal interfaces have obvious benefits for users with special needs who cannot use some or all the communication modes.

A clear disadvantage presented by multimodal interfaces is the special attention needed by the user in coordinating the input modalities, possibly resulting in cognitive overload. Also when receiving multimodal information, the user experiences stimulation of multiple senses, which also affects the cognitive load. From the system-centric view, multimodality requires advanced processes at the combination level and especially at the interpretation level, and also an adaptable approach to presenting information.

The system described in this paper is based on the USIX Interact project (Jokinen, et al. 2002) which aimed at studying methods and techniques for rich dialogue modelling and natural language interaction. In this follow-up project, the main goal of research is to integrate a PDA-based graphical point-and-click interface with the user's speech input, and to allow the system to output in speech as well as drawing on the map. Besides the technical challenges, an important goal is also to investigate possibilities for natural interaction in a route navigation task where the system is to give helpful information about the route and public transportation.

3. Multimodal interaction with MUMS

3.1. User interface

The system can perform two tasks: provide timetable information for public transportation and provide navigation instructions for the user to get from a departure place to a destination. The client application accepts speech and tactile input, and presents information via speech and graphical map data. The touch-screen map interprets all tactile input as locations, so a tap on the screen denotes a pinpoint coordinate location, whereas a circled area will be interpreted as a number of possible locations. The map can be freely scrolled and zoomed in real time, and the inputs are recorded simultaneously and time stamped for later modality fusion phase processing.

In order for the system to be able to retrieve route information, at least the departure and arrival locations must be known, which results in the system returning a route summary containing the most relevant route details. If the user does not provide all necessary information to execute a full route query, the system prompts the user for the missing information. As shown in Example 1 and Figure 1, the user can provide a segment of information either by voice or a map gesture. When all necessary information has been collected, the system will fetch the route details.

Example dialogue 1: The user presents a route query, makes a correction, and finally iterates departure times until a suitable route is found.

U: Uh, how do I get from the Railway station ... uh.

S: *Where was it you wanted to go?*

U: Well, there! + <map gesture>

S: *Tram 3B leaves Railway Station at 14:40, there is one change. Arrival time at Brahe Street 7 is 14:57.*

U: When does the next one go?

S: *Bus 23 leaves Railway Station at 14:43, there are no changes. Arrival time at Brahe Street 7 is 15:02.*

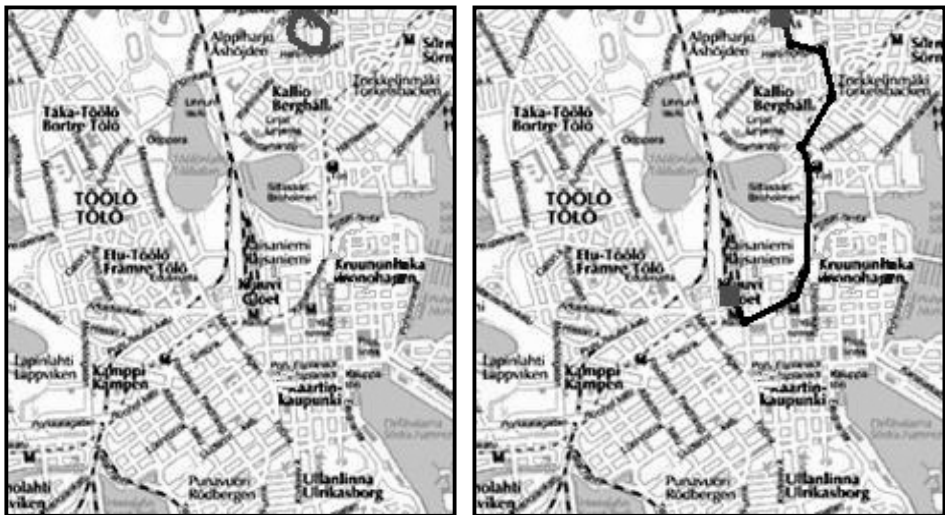


Figure 1. Tactile input (at left) and a graphical representation of a route (at right).

During navigation the route is presented on the screen (Figure 1, at right), and details are supplied by speech, as shown in Example 2.

Example dialogue 2: The user accepts the route suggestion and asks the system to instruct him/her on that route.

U: Ok. Navigate.

S: *Take bus 23 at the Railway Station at 14:43.*

U: Navigate more.

S: *Get off the bus at 14:49 at the Brahe Street stop.*

U: Navigate more.

S: *Walk 200 meters in the direction of the bus route. You are at Brahe Street 7.*

U: Okay.

In addition to requesting route guidance, users can also present questions about route details: travel times, distances, the stop count, etc. Users are also not restricted to any specific timing or form of input, and they can also make corrections to already submitted input at any dialogue phase.

3.2. Interaction model

Coherent frameworks for multimodal interaction patterns are yet to be formed, and thus application development is guided by a few selected studies (e.g. Oviatt et al. 2004), and also by work in the field of linguistics. The purpose of multimodal interfaces is to provide users with a more natural way to interact with a system. In practice, due to the deficiencies in the robustness of natural speech recognition, it is of utmost importance to design the interaction model so that the user is limited to a certain amount of possible ways of presenting information, but at the same time feels that he/she still is free to present this input in a natural and flexible way. The possibility to choose an input strategy is one of the most important factors accounting to naturalness during interaction (Oviatt 2001). The cognitive load experienced by a user is one way to measure an aspect of naturalness. Cognitive load increases e.g. in situations where the user must perform multiple simultaneous tasks or formulate complex utterances.

In MUMS, because of the rather limited functionality and task-specific nature of the system, the user is already limited to a handful of ways of forming a spoken route enquiry, which simplifies the recognition and interpretation processes. The first turn in the dialogue is initiated by the user, who is expected to present a route query. Even though there is just a handful of possible ways of formulating a query, this is clearly the critical point in the dialogue from the performance point of view. In case the user's utterance was misinterpreted or its contents insufficient, the system prompts the user for the missing or additional information one concept at a time in such a way that the user's cognitive load does not affect his/her output.

Several current multimodal applications trade naturalness for robustness by using explicit confirmations. A reliable way to confirm simple details, this approach is however, as Boyce (1999) points out, usually found inflexible and annoying. In the MUMS system, all confirmations are carried out in an implicit manner, as shown in Example 1. This approach is expected to be quite successful here, since the task at hand contains only a few variables. The route summary, presented by speech and graphics, is a straightforward and quick way to determine if the system interpreted his/her input correctly. The next user utterance could e.g. be: *"No, I wanted to get to the opera house"*.

Natural multimodal navigation is a field of research that has resulted in several practical applications. Map representations are a natural way of presenting spatial information, and speech can be used for guidance when the user's eyes are occupied with a mobile task. The success of a guidance task depends also on the naturalness and the cognitive load experienced by the user. The cognitive load can be reduced e.g. by presenting the information in suitable chunks (Cheng et al. 2004). Also important are the way the information is split into the output modalities, and the used level of detail.

In our approach graphical output is at the moment simple and static; a fetched route is kept on the map as long as the user is en route. Users can however choose the detail of the spoken instructions. The default detail level is intended for experienced commuters, and consists of just the basic details, e.g. times and locations. In the detailed level, information is presented in smaller chunks and the user is provided with additional details, such as the number of stops, travel times, etc. The detailed level is aimed at users needing clearer instructions, such as the visually impaired. At the moment the navigation level is set by the user, but we can also envisage that it would be possible for the system to adapt itself, by its knowledge of the particular situation and learning through interaction with the user, when to switch to a more detailed navigation mode.

4. Conclusion

The presented MUMS system provides the user with a natural way to enquire locational information and route navigation. The system's interaction level is designed for ease of use and naturalness, keeping in mind the challenges set by especially the current state of speech recognition. We assume that the system architecture is general enough to be used in other similar multimodal applications as well.

Further studies will aim at improving the integration and synchronisation of information in multimodal dialogues. The system will also be extended to handle more complex pen gestures, such as areas, lines and arrows. As the complexity of input increases, so does the task of disambiguation of gestures with speech, which will undoubtedly present us with new challenges. Usability testing of the prototype system will be started soon.

5. References

- Boyce, S. 1999. Spoken natural language dialogue systems: User interface issues for the future. In: Gardner-Bonneau (ed.). *Human Factors and Voice Interactive Systems*. 37—62.
- Cheng, H.; Cavedon, L.; Dale, R. 2004. Generating Navigation Information Based on the Driver's Route Knowledge. In: Gambäck B.; Jokinen, K. (eds.) *Procs of the DUMAS Final Workshop Robust and Adaptive Information Processing for Mobile Speech Interfaces, COLING-2004 Satellite Work-shop*, Geneva, Switzerland. 31—38.
- Gibbon, D.; Mertins, I.; Moore, R. 2000 (eds.). *Handbook of Multimodal and Spoken Dialogue Systems. Resources, Terminology, and Product Evaluation*. Dordrecht: Kluwer.
- Jokinen, Kristiina; Kerminen, A.; Kaipainen, M.; Jauhiainen, T.; Wilcock, G.; Turunen, M.; Hakulinen, J.; Kuusisto, J.; Lagus, K. 2002. Adaptive Dialogue Systems – Interaction with Interact. In: *Proceedings of the 3rd SIGdial Workshop on Discourse and Dialogue*. Philadelphia, USA: Association for Computational Linguistics.
- Jokinen, Kristiina; Raike, Antti 2003. Multimodality – technology, visions and demands for the future. In: *Proceedings of the 1st Nordic Symposium on Multimodal Interfaces*. Copenhagen.
- Oviatt, Sharon; Cohen, P.R.; Wu, L.; Vergo, J.; Duncan, L.; Suhm, B.; Bers, J.; Holzman, T.; Winograd, T.; Landay, J.; Larson, J.; Ferro, D. 2000. Designing the User Interface for Multimodal Speech and Pen-based Gesture Applications: State-of-the-Art Systems and Future Research Directions. In: *Human Computer Interaction*, 15(4). 263—322.
- Oviatt, Sharon 2001. Advances in Robust Processing of Multimodal Speech and Pen Systems. In: Yuen, P.C. and Yan, T.Y. (eds.) *Multimodal Interfaces for Human Machine Communication*. London, UK: World Scientific Publisher.
- Oviatt, Sharon; Coulston, R.; Lunsford, R. 2004. When Do We Interact Multimodally? Cognitive Load and Multimodal Communication Patterns. In: *Proceedings of the Sixth International Conference on Multimodal Interfaces (ICMI 2004)*, Pennsylvania, USA. 14—15.