

Mobile Speech-based and Multimodal Public Transport Information Services

Markku Turunen, Topi Hurtig, Jaakko Hakulinen

University of Tampere, Department of Computer Sciences

{firstname.lastname}@cs.uta.fi

Ari Virtanen, Sami Koskinen

VTT Technical Research Centre of Finland

{firstname.lastname}@vtt.fi

ABSTRACT

Mobile devices, such as smartphones and personal digital assistants, can be used to implement efficient speech-based and multimodal interfaces. In this paper we present three approaches for mobile public transport information services, such as bus timetables and route guidance. These applications offer varying functionality depending on the devices and modalities used. Some services can be used with regular mobile phones, while other services use GPS information and multimodal inputs to make the interaction more efficient. The applications are designed with various user groups, such as visually impaired users, in mind, and evaluated with long-term pilot studies and usability tests.

Categories and Subject Descriptors

H.5.2 [Information Systems]: User Interfaces - *voice I/O, natural language, interaction styles, prototyping evaluation.*

D.2.11 [Software Engineering]: Software Architectures - *Domain-specific architectures.*

General Terms

Human Factors.

Keywords

Speech interfaces, mobile interaction, multimodality.

1. INTRODUCTION

The interest towards mobile spoken and multimodal dialogue applications has increased constantly. Most of the research has concentrated on the use of speech to circumvent problems of small displays in graphical applications [3]. On the other hand, we can increase the usability of speech-based applications with the optimal use of additional resources, such as supporting modalities (e.g., graphics and haptics) and information sources (e.g., positioning information) when available.

The use of speech in mobile user interfaces is no longer limited to standard telephony use. New devices, such as smartphones, can be used in more versatile ways. Different devices have different features that can be employed in interaction. Items such as graphical menus and commands that are always available can be easily presented graphically, while this is a challenging task in a

speech only environment. In general, mobile device resources can be used to support speech interfaces when available.

Multimodality can bring many advantages to speech-based interaction. Especially concerning spatial information, e.g. giving navigational instructions in guidance applications, spoken verbal explanations may not be the most effective way of exchanging information for normally sighted people [9]. Instead, interactive maps can be very efficient for displaying navigation information. Similarly, touch screens can be used for tactile inputs, and GPS (Global Positioning System) information reduces the need to enter position information (e.g., departure places) manually, and makes more sophisticated services, such as route guidance, possible. On the other hand, spoken guidance is the only possibility for visually impaired people, so it would be beneficial to offer a multimodal interface to support different users.

In the Finnish project "New Methods and Applications of Speech Technology" (PUMS) mobile speech-based interaction is studied from technology and human-computer interaction perspectives. The main issues addressed in this paper are: (1) distributed service infrastructure and dialogue management, (2) route guidance and navigation, (3) multimodality, and (4) special user groups, such as visually impaired users.

These issues are addressed in three public transport information services called Stopman, NOPPA, and MUMS. The domain was selected because public transport information services are particularly suitable for mobile interaction, and they are highly relevant for large user populations. In particular, speech-based mobile applications can be the only possibility for such special user groups as visually impaired people to access the needed information when using public transport.

The applications have different characteristics based on their focus areas. Stopman offers information on bus timetables, while the other applications offer route guidance and navigation information. Functionalities of the systems will be presented in more detail in the following sections with concrete example dialogues. There are differences in the implementation platforms as well, as illustrated in Figure 1. There are two versions of Stopman, the original telephone version, and the multimodal smartphone version. Similarly, NOPPA was originally PDA-based, but there is a smartphone version as well. The properties of the smartphone versions are marked with (*). The MUMS system operates on a

PDA. Dialogue management takes place in the server (MUMS, telephone-Stopman), in the terminal device (NOPPA), or it is distributed (smartphone-Stopman).

	Telephone	PDA	Smartphone	Server DM	Terminal DM
Stopman	●		● (*)	●	● (*)
NOPPA		●	● (*)		●
MUMS		●		●	

Figure 1: Terminal devices and dialogue management.

The modalities used by the applications are illustrated in Figure 2. All systems have a speech interface, i.e., speech inputs and outputs, and Stopman and NOPPA use additional (telephone) keys for interaction. The smartphone version of Stopman has a graphical user interface, and MUMS combines a graphical map with tactile inputs. The guidance function of NOPPA relies on the use of GPS information, while MUMS includes an option for it.

	Speech	DTMF/KEY	GUI	Map+Tactile	GPS
Stopman	●	●	● (*)		
NOPPA	●	●			●
MUMS	●			●	(optional)

Figure 2: Modalities used in the applications.

Next, we present the general server infrastructure, and the common server architecture. Then we present the different applications, and results of their evaluations.

2. SERVICE INFRASTRUCTURE

Most of the current mobile speech systems are distributed in a way that part of the functionality takes place in a mobile terminal device, while some other functionality and information sources are located in one or more servers. There are many reasons for this. Most importantly, most mobile devices are not able to run complete multimodal spoken dialogue systems because of insufficient resources and missing technology components (e.g., speech recognition). Second, many applications, and information services in particular, need external information sources, such as continuously updated databases. Third, more and more applications are modeled using the Service-Oriented Architecture (SOA) approach, where loosely coupled and interoperable services are running on networked environments. For these reasons, most mobile speech systems consist of at least one terminal, one server, one database, and various other services. In addition, there is an increasing trend to use positioning information in mobile services. This general infrastructure is illustrated in Figure 3.

As illustrated in Figure 3, there are many ways how the resources in the infrastructure can be used. For example, “Device 1” has local speech synthesizer, while “Device 2” has GPS capabilities, and “Device 3” has local speech recognizer. The available server-based services include speech recognition, speech synthesis, and database services. Furthermore, most devices and servers are capable of running dialogue management tasks, and in this way truly distributed dialogues can take place. This will be further discussed in the next section.

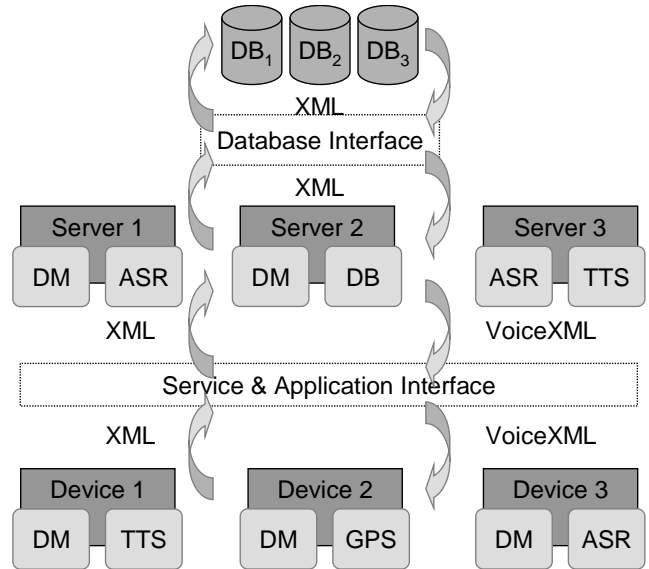


Figure 3: Mobile service infrastructure.

In the architecture illustrated in Figure 3, XML-based interfaces are used between the different infrastructure components, and VoiceXML is used to exchange dialogue information between the servers and the terminal devices. In addition, both interfaces (database and service) can act as proxies to cache commonly used information, such as public transport timetables and routes. The three mobile applications presented in the rest of the paper are variants of this generic model. In all applications, speech recognition is performed by the server using the same speaker independent speech recognizer. Otherwise, the applications vary on the use of resources. Next, we present the common system architecture used in the timetable and multimodal route guidance applications.

2.1 Server System Architecture

The timetable and multimodal route guidance systems are built on top of the Jaspis architecture [14]. Jaspis is based on the agents – managers – evaluators – paradigm, where managers coordinate sets of evaluators to select appropriate agents that actually handle the tasks (e.g., dialogue decisions). Agents in Jaspis-based applications are compact software components. Managers represent the high level organization of tasks that are handled within spoken dialogue applications. All components share all the information via Information Storage.

Figure 4 illustrates the components of the server system. The systems contain standard Jaspis components (Communication, Dialogue, Presentation and Input managers), and an additional manager, the Database Manager, which is used to communicate with the databases. The differences in the systems are in the structure of the Information Storage (modeling of dialogue and domain information), and in the agents that use that information to carry out interaction tasks. Next we present the mobile version of the Jaspis architecture targeted for distributed mobile speech-based and multimodal dialogue systems.

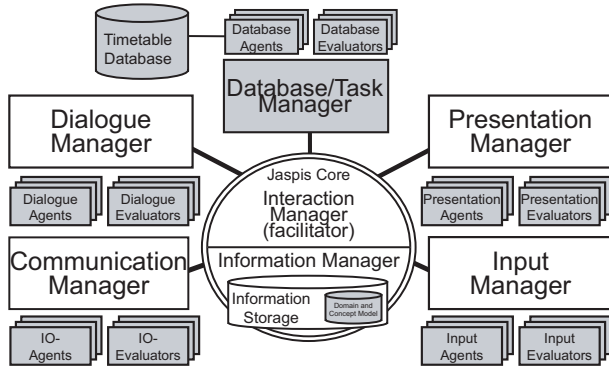


Figure 4: Server system architecture.

2.2 Mobile Architecture for Terminal Devices

We have extended the Jaspis architecture to support mobile spoken dialogue systems [16]. The mobile version implements the core architecture for mobile devices running J2ME (Micro Edition of the Java 2 Platform) environments, such as Series 60 mobile phones. With the unified base architecture and compact software agents we are able to create versatile applications where the tasks can be distributed efficiently between servers and mobile devices. In addition, the mobile architecture is capable of running complete spoken and multimodal dialogue applications in the mobile device. This allows iterative construction of mobile applications starting from the situation in which most of the functionality is performed on the server, but gradually more functionality can be moved from the server to the mobile device in the form of compact agents. Next we present two versions of the Stopman timetable system. The speech-only version is fully server-based, and is accessed with mobile phones. The multimodal version of the system is distributed between a server and a smartphone.

3. SPEECH-BASED TIMETABLES

We have developed multiple spoken and multimodal timetable systems on top of the Jaspis system architecture in previous projects. The results suggest that an open, user-initiative dialogue strategy based on the data collected from human-human interaction fails to provide robust enough interface [15]. Other research has shown similar results [12]. In the case of the Stopman system we developed a task-oriented interface that provides the basic functionality in a system-initiative manner, while the rest of the functionality is available with a user-initiative interface.

The Stopman system provides timetables for each of the about 1200 bus stops in Tampere City area. First, the system requests the user to give a bus stop name, as demonstrated in Example 1 (S_1). The aim of the system is to satisfy most of the callers with the first timetable listing (S_3). After this, the rest of the functions are available. Functionality includes navigation in the timetable, selection of specific bus lines, and specifying a certain time. For example, the user may request timetables for the bus line 30 passing the main library on Saturday after 16.25.

The Stopman system has been publicly available since August 2003, and multiple usability evaluations have been conducted to the system to make it more efficient and pleasant to use. Concluding from the usage experiences, long listings and spoken menus were frequent problems. Users found them boring and compli-

cated, and they easily forgot what was said previously. Results of the evaluation are discussed in more detail in Section 6. Next we present how these problems can be addressed with a multimodal distributed interface to be used with smartphones.

S_1 :	Welcome to Stopman. You can ask help by saying "tell instructions". Please tell the name of the bus stop, for example "Central Square".
U_1 :	"Alexander Church".
S_2 :	Do you want to retrieve timetable for stop "Alexander Church"?
U_2 :	"Yes."
S_3 :	The time is 10:10. Today, the next buses leaving from bus stop "Alexander Church" are to "Vehmainen" number 5, now (list continues)... Please give your command, for example "tell instructions".
U_3 :	"Select line."
S_4 :	Please tell the line number. You can list the lines by saying "list lines".
U_4 :	"Thirteen."
S_5 :	Next number thirteen buses from "Alexander Church" are to Hermia in 1 minute, to "Ikuri", in (list continues)...

Example 1: An example dialogue with the Stopman system.

3.1 Multimodal Smartphone Version

In the original telephone system all components are running on the server and the user access the system by calling the system with a regular mobile phone. In the smartphone version the components are distributed between the server and the mobile device. In both versions the server generates high-level dialogue descriptions using VoiceXML. These descriptions are carried out by the smartphone or the server with available resources. In the recent research several solutions are proposed to execute VoiceXML descriptions in devices with limited processing capabilities (e.g., [1] and [11]). In the current version of Stopman the mobile device contains an interpreter for executing VoiceXML descriptions with additional multimodal information. In the future versions some of the components running on the server will be moved to the mobile terminal device.

The mobile version is based on a generic model for distribution of dialogue management tasks to available terminal devices [13]. The role of the server in this model is to handle the high-level dialogue management by generating extended VoiceXML descriptions that provide all the information needed to execute multimodal dialogue tasks (e.g., menus) in the mobile devices. The tasks are realized with available resources, such as speech recognizers, synthesizers and graphical capabilities. This way the resources available in each environment can be used optimally as devices that know their own resources make the decisions how the descriptions are actualized. This is illustrated in Figure 5.

Figure 6 illustrates the two user interfaces of the Stopman application that are utilizing the distributed model. The interaction is similar with both interfaces except that the smartphone interface uses the display to show possible menu options and recaps of system prompts. Even though VoiceXML is originally built for spoken interaction only, it is possible to use it for these kinds of rather simple multimodal tasks.

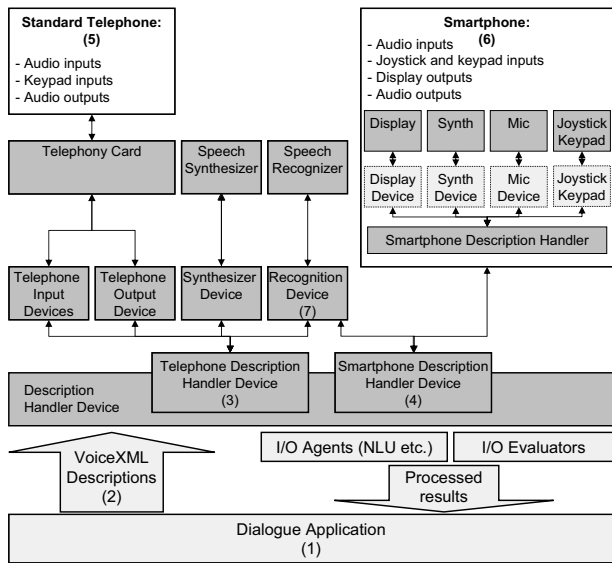


Figure 5: Distribution of dialogue tasks.

A short dialogue with the telephony interface is illustrated on the left hand side of Figure 6. With the telephony interface speech and touch tones can be used to interact with the system. Forms are interactive so that users can use speech, or navigate between options by the telephone keypad. The active item is selected and spoken to the user.

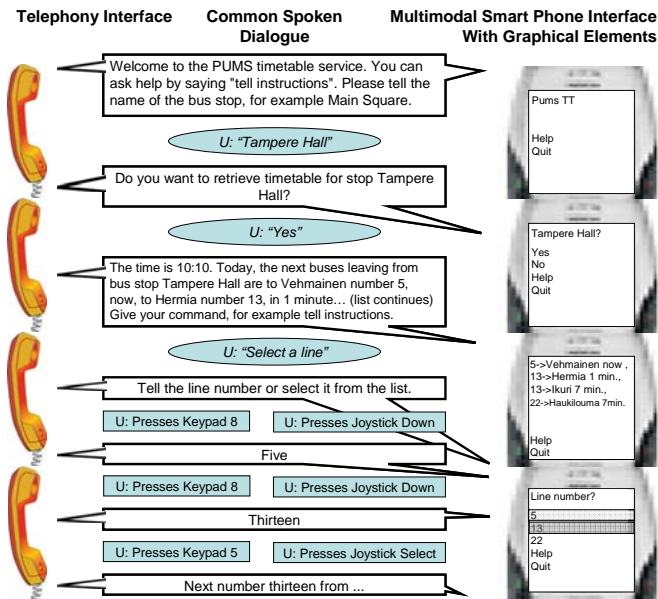


Figure 6: Multimodal example dialogue.

The smartphone interface is illustrated on the right hand side of Figure 6. In this interface display, joystick and keypad are used as additional modalities. In addition to spoken prompts, supporting textual information is presented on the screen, and menus are presented graphically. Items can be selected by using the joystick or the keypad like in the telephony version. When an option is selected it is highlighted and its value is spoken.

As illustrated in Figure 6, the spoken prompts and the displayed textual contents are not the same on surface level. It is more useful to show only the essential information on the screen whereas more words are needed in order to make the prompt understandable and speech synthesis fluent. The conveyed information is, however, the same. The use of the display could be made even more efficient by utilizing other forms of graphics, such as maps. This approach is studied in more detail in the multimodal route navigation application. Next we present the speech-based route navigation and guidance application.

4. ROUTE NAVIGATION AND GUIDANCE

Speech-based route navigation is studied with the NOPPA application. The system offers services such as route guidance and service disruption information (e.g. roadworks information). The system is meant to be an information aid for visually impaired people in everyday life, and it contains several other services and features in addition to public transport information services [18].

The NOPPA architecture is presented in Figure 7. Due to the low processing capacity of mobile terminals and the rather low bandwidth of wireless data connections, most of the work is done in the Information Server. Data flow between the mobile terminal and the Information Server is minimized, which decreases communication costs and shortens response times.

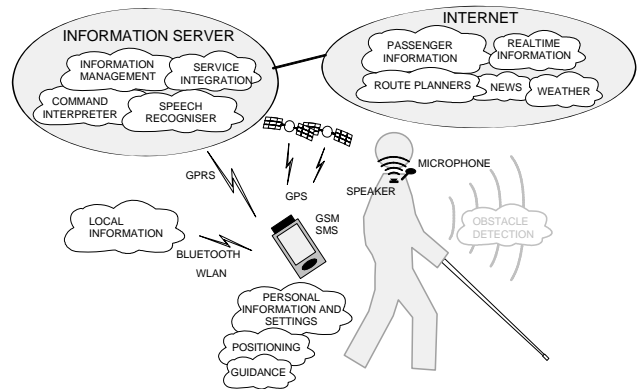


Figure 7: Architecture of the NOPPA guidance system.

Next we present how the guidance information will be gathered and represented to the user with a speech interface.

4.1 Unbroken Trip Chain

If we examine problems a visually impaired person meets when using public transport, we recognize the following list (depends slightly on the means of transportation):

- trip planning* - finding a stop/station - finding an entrance to the station - navigating inside the station - finding the right platform and waiting place - knowing when the right vehicle arrives - finding a vehicle entrance - payment - finding a seat - depart on right stop - navigating inside the station - finding the exit of the station - finding the destination

Most of these tasks are trivial for the sighted, but very difficult for the visually impaired. There are cases when a blind person has spent several hours on the bus stop, because he/she could not recognize the arrival of the specific vehicle.

To provide an unbroken trip chain for visually impaired users, we must switch seamlessly between different modes of operation during the trip (see Figure 8). This requires that the system must be context-aware to recognize transition points and accordingly automatically change its mode of operation.

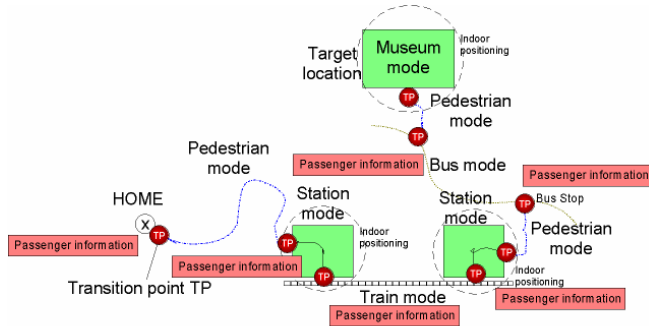


Figure 8: Unbroken trip chain and different operation modes.

An example of the information chain described above is shown in Figure 9. The route plan has the information of street names, used vehicles, line numbers, stop names, coordinates and times. The main difference to car navigation is the time information. During the journey we have to reach the bus stop before our bus passes it.

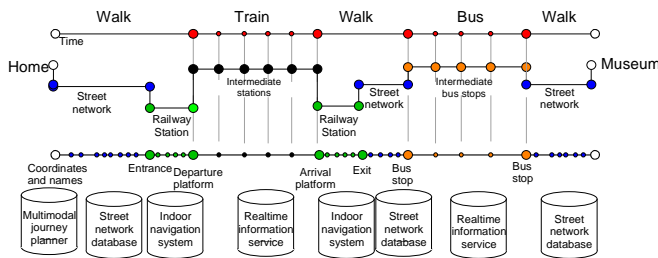


Figure 9: An example with a chain of unbroken information.

With positioning one can follow the route and pick up relevant information from different passenger information systems. Real time information, as well as disturbance information, comes typically from different sources than the journey plan. The route plan is augmented with road work information, Point of interest (POI)-information and Area of interest (AOI)-information.

The operating mode of the guidance system can be changed automatically according to current location and time. First, the system guides the user to the bus stop, then changes its mode and starts to follow the real time arrival time of the bus. Inside the bus the system starts to follow bus stops and instructs when to leave the bus, and finally guides the user to the destination. This all can be done without user intervention or action. Real time information plays an important role through the unbroken trip chain [6]. Integrating indoor navigation to outdoor navigation is also a great challenge [19].

4.2 Guidance with Speech

Spoken guidance information is very challenging for many reasons. In the pedestrian mode we have to navigate in street networks. Current maps used in navigation systems are made to car navigation purposes. Coordinates describe the street centerline and every route goes middle of the road. This is not adequate for

pedestrians, because pedestrians use pavements, zebra crossings and footpaths. Correct routes are crucial for the blind persons, because they can not see if the route proposed by the navigator is dangerous or unusable. Because of these inaccuracies and inappropriate maps, guidance can not be the same as used in car navigation. These systems can use simple "turn right" or "turn left" instructions.

The guidance method used in NOPPA takes account of possible inaccuracies and can change its operation depending on positioning accuracy. Guidance dialogues imitate the way people act when giving instructions to other people. Because speech dialogues are distributed over several processes, and between the server and terminal, we have to solve priority problems. Some of the dialogues are merely informative, while the others need to be spoken at the right time. Important messages should be capable of overriding less important information. One interesting research topic in the future will be the use of different voices to place emphasis on important messages.

4.3 Interaction with the NOPPA System

In the beginning of the interaction with the NOPPA system the user selects system functions with keys, as demonstrated in Example 2 (U_1, U_3). If the departure or the destination address is unknown, the system requests the address (S_3). The speech recognition grammar contains 15,000 Helsinki area street names. GPS positioning makes planning easier, because the system knows the current address by using reverse geocoding, and often used target addresses can be bookmarked. After that, the user listens to the suggested route plan (S_4), and if he/she considers it acceptable, the next step is to start route following and guidance (S_5). Otherwise, he/she can re-plan the route.

U_1 :	<selects "train timetables" with PDA keys>
S_1 :	Train timetables. Say station of departure, destination and time
U_2 :	Tampere, Helsinki, now
S_2 :	Recognition ... Tampere Helsinki now, confidence 439. Retrieving trains... 17:07 Tampere 18:56 Helsinki Intercity two 180 Helsinki no changes duration 1:49. Next departure 17:36 Tampere...
U_3 :	<selects "journey planning" with PDA keys>
S_3 :	Please tell the destination place or address, or select it from the list.
U_4 :	Koskikeskus
S_4 :	Recognized "Koskikeskus", confidence 707. Retrieving route from journey planner ... route ready. Starting 0 min, no changes. Walk from Tekniikkankatu 2 19:09 to Poliisikoulu 19:14 Northwest 0.3 km. Bus line 13 from Poliisikoulu 16:57 Railroad station 19:33. Walk from Railroad station 19:33 Koskikeskus 19:35 South 0.5 km.
S_5 :	<after that the route guidance starts>

Example 2: An example dialogue with the NOPPA system.

A challenging task is to produce spoken descriptions of the candidate routes to enable the visually impaired users to create a proper mental image of the route and its usefulness. Map representation is in this case superior, one can see at once if the route is sensible or not. This is illustrated in the following section in the case of multimodal route navigation.

5. MULTIMODAL ROUTE NAVIGATION

In the MUMS system we study the possibilities of rich multimodality in speech-based public transport information services. The system offers similar information as the NOPPA system. In addition to speech inputs and outputs it allows pen-pointing gestures on a graphical map. Speech and pen input are known to be very closely coupled, and their combined use has been extensively studied. Studies conducted with the QuickSet system [10] show 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. Users are free to use any chosen combination of input and output modalities, resulting in natural, flexible and especially efficient task-based interaction. In addition to the touch screen, additional location data is provided by GPS information. This data can also be used for assisting the system in parsing route navigation instructions.

5.1 System Overview

The user interface is implemented into a mobile terminal device. At the moment the client application is implemented into a PDA, but basically any mobile device with touch screen, and sound and network I/O capabilities could be used. The client device uses the GPRS network to communicate with the system server, which in turn accesses an external route database. The system server handles all processing of the user-provided information, and, apart from a lightweight speech synthesizer, the client can be considered only a simple user interface. The server software is built on top of the Jaspis framework (see Section 2.1). Thanks to its configurability, it has been modified for the use of multiple modalities. The external routing system and database returns, for each complete query, a detailed set of route information in XML format. A more detailed system description can be found in [5].

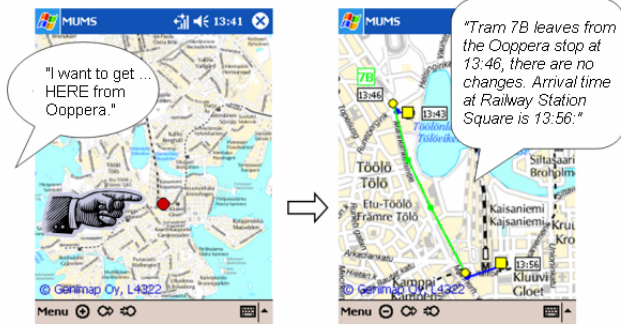


Figure 10: Multimodal interaction.

5.2 User Interface

Users can freely roam and zoom the map in real-time. User input - speech and map gestures - is simultaneously recorded with the help of a recording button. A full sample dialogue is shown in Example 3, with Figure 10 illustrating multimodal system usage. Users are free to present input unimodally with either speech or gestures, or multimodally with a combination of both (U_2 in Example 3). Figure 10 shows a system response, where synthetic speech is combined with a graphical representation. When a user accepts the route the system proposes (S_3), the navigation phase is

initiated by saying "navigate" or "start navigation" (U_4). The user is then guided along the route a leg at a time, in the normal (S_4, S_5, S_6) or the detailed guidance level.

U_1 :	Uh, how do I get from Ooppera to ... uhh?
S_1 :	Where was it you wanted to go?
U_2 :	Well, there! + <map gesture>
S_2 :	Tram 7B leaves from the Ooppera stop at 13:46, there are no changes. Arrival time at Railway Station Square is 13:56.
U_3 :	When does the next one go?
S_3 :	Bus 43 leaves from the Ooppera stop at 13:48, there are no changes. Arrival time at Railway Station Square is 13:59.
U_4 :	Ok. Navigate.
S_4 :	Take bus 43 from the Ooppera stop at 13:48.
U_5 :	Navigate more
S_5 :	Get off the bus at 13:53 at the Elielinaukio stop.
U_6 :	Navigate more.
S_6 :	Walk 150 meters; you are at the Railway Station Square.

Example 3: An example dialogue with the MUMS system.

5.3 Dialogue Management and Fusion

The processing of received user input begins with the recognition of each modality and the attaching of high-level task-relevant concepts, e.g. "explicit_speech_location" to input units. The next phase, the fusion of modalities results in an N-best list of user input candidates. The process then advances to the dialogue management module, which, while being able to access the dialogue history, attempts to determine user intentions, chooses the input candidate that best fits the situation and task at hand, and carries out the corresponding task. Depending on the content of the user input and the state of the dialogue, the dialogue management module forms a generic response, which is then accessed by the presentation module. The presentation module formats the response according to the set user preferences and the client hardware in use, after which the information is ready to be sent and presented in the terminal device. Detailed information about the system's dialogue model and fusion mechanisms can be found in [4].

6. EVALUATION AND USER STUDIES

All three applications have been tested with real users. NOPPA has been used by visually impaired pilot users over two years to guide the development. Furthermore, the content of the services offered to the users was evaluated with ten users with varying degrees of visual impairments. Stopman and MUMS have been evaluated with usability tests, and Stopman has been in public use for more than 30 months. To conclude, based on subjective evaluation and user feedback, users have been very satisfied with the systems. Next we report the main findings from the evaluation and user studies of the timetable and multimodal route navigation systems.

6.1 Pilot Use of the Timetable System

The Stopman system has been publicly available since August 2003. All calls to the system are recorded and logfiles have been analyzed. The first version was in public use for fifteen months

from November 2003 to January 2005. The second version of the system has been in public use since February 2005. In addition, various usability evaluations (e.g., subjective evaluations) have been conducted to the system to make it more efficient and pleasant to use. In a recent study [17] we analyzed 1855 dialogues from a 30 month period, and compared how the interaction differs between usability studies, first months of the real use, rest of the real use, and between the different versions of the system.

The comparison of different use categories shows that the results obtained with usability tests differ significantly from those gained from real usage, and the data from initial use differs significantly from the data collected after that. There are highly significant differences between the first month and the rest of the pilot usage in almost all aspects of the system use (e.g., help requests, interruptions, speech recognition rejections, silence timeouts, and repeat requests). The differences between the system versions were quite few, while the differences between real use and usability studies were extremely high in almost all aspects. For example, the call was ended explicitly in 65% of the dialogues during the usability studies, while in the real use only 5% of calls were such. Similar results were obtained on the use of different modalities (speech, touch stones), different system functions etc.

6.2 Multimodal Interaction Studies

In a recent evaluation [1] we focused on users' experience on the multimodal aspects of the MUMS system. Especially, we set to study the users' preference over speech or the tactile interface, if the users' age and gender influenced their expectations, and how the users' expectations differed from their actual experience with the system. The subjective evaluation utilizes in parts a modified version of SERVQUAL that has already proved to be informative in subjective evaluation of dialogue systems [2].

The test included 17 test users, ten male and seven female, aged between 23 and 61, of different professions. When studying user expectations in different age groups, we find that users from ages 33 to 48 had very low expectations for system performance and the usability of the system's multimodal aspects. Interestingly, when we compared expectations with actual performance evaluations, the scores for this middle group rose to the same level with the younger and older groups, i.e. their actual experience was very positive. When divided into two age groups, younger people seem more willing to use the system unimodally, while older people seem to exploit the multimodal aspects of the system.

Gender seems to affect the differences between expectations and perceived qualities only in a few cases; most of them concern the system's empathic capabilities. This phenomenon has previously been noted also in speech-only systems [2]: female users consider the system as more understanding and considerate than what they expect. The female group was also surprised by the system's multimodal usability, considered the system easy to adopt and asserted strong enthusiasm about using the system in the future.

Users were divided into two groups based on information given to them before the evaluation: the speech group was trained and instructed to interact with a speech interface that has a tactile input option, while the tactile group was trained and instructed to interact with a tactile system that has additional spoken dialogue capabilities. Interestingly, the priming effect did play a role here: the speech group seemed more willing to use a tactile interface

unimodally than the tactile group, and was disappointed with the use of a speech only system. Analogously, the tactile group felt slightly more positive about using speech input and output, than what they expected. We also observed that the tactile group was happier with the system's performance, especially with the rate of how often the system interprets user input correctly. And, interestingly, the speech group seemed to feel, more than the tactile group, that the system is slow, even though the response time of the system is not affected by the form of user input.

7. CONCLUSIONS AND FUTURE WORK

In this paper we have presented three approaches for speech-based and multimodal mobile public transport information systems. In order to address the challenges of distributed heterogeneous service infrastructures, we have presented an architecture that enables the construction of mobile multimodal distributed dialogue systems, and a generic model for distributing dialogue tasks between servers and mobile terminals. We presented a multimodal mobile version of an existing server-based speech application. Our future work focuses on more complex mobile situations, where dialogues can be temporally discontinuous, and devices used for communication may change during the dialogue. This brings great challenges for modelling and implementation of distributed dialogues.

In the case of speech-based route guidance, we introduced a solution that makes the use of public transportation very comfortable for visually disabled users. Moreover, all users benefit from the possibility to keep the mobile terminal device tucked away during travel while receiving personal travel information via speech. In the future our aim is to add support for routine travel, implement user selectable guidance levels, and add speech comments to personal recorded routes. One great challenge is to improve speech recognition capabilities, which is a solid requirement for nationwide operation. Also, a lot of work has to be done to get necessary information available in general. From the user perspective, valuable feedback and ideas for new functions are continuously provided by pilot users.

In the area of multimodal route navigation, we presented solutions that allow users freedom in the choice of used input and output modalities, thereby making interaction with the system possible for different users and various situations. However, one must not forget that the use of multiple modalities is not an all-around solution to mobile interaction; multimodality seems generally to improve performance, but mainly in spatial domains, such as map and navigation applications. In addition, users are obviously not ready to use a complex multimodal system – without training they don't know what to do, as is pointed out also in recent studies [8].

In the user evaluations we have noted the importance of ecologically valid evaluation settings, which affect the interaction considerably. According to our studies, the usability test data is more similar to data from initial usage than that of later months. Still, there are highly significant differences between the usability tests and the initial usage. The results suggest that when we are dealing with mobile speech applications, and emerging technologies in general, ecological validity plays crucial role in the evaluation. In order to address this challenge, we should evaluate the applications in real context, and deliver prototypes to public use in the early stages of the development. In addition, there is a need for new evaluation methods and subjective methods in particular.

The user evaluations pointed out that individual users clearly regard practical speech-based interactive systems differently. Although the differences in evaluation cannot always be traced to stem from prior knowledge, predisposition, age, or gender differences, it is important to notice that the goal of building one single practical system that would suit most users is not reasonable. Rather, there is a need for adaptivity and personalization, allowing the use of different input strategies, with responses tailored according to user preferences. As use of interactive systems in controlled evaluation settings is known to differ from actual use, we are planning a more extensive evaluation, in which system usage is monitored over a period of several days in everyday commuting situations. In addition to collecting new data, this will also reduce the effect of subjective novelty-related issues.

In the current applications we have focused on specific areas of speech-based and multimodal interaction in mobile settings. Still, the applications share common resources in many cases. For example, all systems are variants of the same generic service infrastructure, and they use the same speech technology components and information sources. Furthermore, the timetable and multimodal route navigation applications use the same server system architecture for task such as dialogue management, while the speech-based route guidance application handles dialogue management in the mobile terminal device. In the future we focus on the tighter integration of the services. Our ultimate goal is to offer a wide range of services with a unified adaptive multimodal speech interface that can be accessed with different mobile terminal devices ranging from regular mobile phones to full-featured PDAs. For example, we will continue the development of the multimodal route navigation system by implementing a smart-phone version of the terminal application.

8. ACKNOWLEDGMENTS

This work is supported by the Technology Development Agency of Finland (Tekes), PUMS-project.

9. REFERENCES

- [1] Bühler, D., and Hamerich S.W. Towards VoiceXML Compilation for Portable Embedded Applications in Ubiquitous Environments. In *Proceedings of Interspeech 2005*: 3397-3400, 2005.
- [2] Hartikainen, M., Salonen, E.-P., and Turunen, M. Subjective Evaluation of Spoken Dialogue Systems Using SERVQUAL Method. In *Proceedings of ICSLP 2004*: 2273-2276, 2004.
- [3] Hemsén, H. Designing a Multimodal Dialogue System for Mobile Phones. In *Proceedings of Nordic Symposium on Multimodal Communications*, 2003.
- [4] Hurtig, T., and Jokinen, K. Modality Fusion in a Route Navigation System. In *Proceedings of the Workshop on Effective Multimodal Dialogue Interfaces EMMDI-2006*, 2006.
- [5] Hurtig, T., and Jokinen, K. On Multimodal Route Navigation in PDAs. In *Proceedings of 2nd Baltic Conference on Human Language Technologies HLT'2005*, 2005.
- [6] Jokinen, K., Hurtig, T. User Expectations and Real Experience on a Multimodal Interactive System. In *Proceedings of Interspeech 2006*, 2006.
- [7] Koskinen, S., and Virtanen, A. Public transport real time information in Personal navigation systems for special user groups. In *Proceedings of 11th World Congress on ITS*, 2004.
- [8] Krüger, A., Butz, A., Müller, C., Stahl, C., Wasinger, R., Steinberg, K-E., and Dirschl, A. The Connected User Interface: Realizing a Personal Situated Navigation System. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI 04)*, 2004.
- [9] Oviatt, S. Multimodal interfaces for dynamic interactive maps. In *Proceedings of the SIGCHI conference on Human factors in computing systems: common ground, CHI '96*: 95-102, 1996.
- [10] Oviatt, S., Cohen, P.R., Wu, L., Vergo, J., Duncan, L., Suhm, B., Bers, J., Holzman, T., Winograd, T., Landay, J., Larson, J., and Ferro, D. Designing the User Interface for Multimodal Speech and Pen-based Gesture Applications: State-of-the-Art Systems and Future Research Directions. *Human Computer Interaction*, 15(4): 263-322, 2000.
- [11] Rajput, N., Nanavati, A. A., Kumar, A., and Chaudhary, N. Adapting Dialog Call-flows for Pervasive Devices. In *Proceedings of Interspeech 2005*: 3413- 3416, 2005.
- [12] Raux, A., Langner, B., Bohus, D., Black, A. W., and Eskenazi, M. Let's Go Public! Taking a Spoken Dialogue System to the Real World. In *Proceedings of Interspeech 2005*: 885-888, 2005.
- [13] Salonen, E. -P., Turunen, M., Hakulinen, J., Helin, L., Prusi, P., and Kainulainen, A. Distributed Dialogue Management for Smart Terminal Devices. In *Proceedings of Interspeech 2005*: 849-852, 2005.
- [14] Turunen, M., Hakulinen, J., Rähkä, K.-J., Salonen, E.-P., Kainulainen, A., and Prusi, P. An architecture and applications for speech-based accessibility systems. *IBM Systems Journal*, Vol. 44, No. 3: 485-504, 2005.
- [15] Turunen, M., Hakulinen, J., Salonen, E.-P., Kainulainen, A., and Helin, L. Spoken and Multimodal Bus Timetable Systems: Design, Development and Evaluation. In *Proceedings of 10th International Conference on Speech and Computer (SPECOM 2005)*: 389-392, 2005.
- [16] Turunen, M., Salonen, E.-P., Hakulinen, J., Kanner, J., Kainulainen, A. Mobile Architecture for Distributed Multimodal Dialogues. In *Proceedings of ASIDE 2005*, 2005.
- [17] Turunen, Hakulinen, J., and Kainulainen, A. Evaluation of a Spoken Dialogue System with Usability Tests and Long-term Pilot Studies: Similarities and Differences. In *Proceedings of Interspeech 2006*, 2006.
- [18] Virtanen, A., and Koskinen, S. Information Server Concept for Special User Groups. In *Proceedings of the 11th World Congress on ITS 2004*, 2004.
- [19] Virtanen, A., and Koskinen S., Towards Seamless Navigation. In *Proceedings of the MobileVenue'04*, 2004.