Preface

This thesis concludes my Master's of Science degree in Communication Technology at Norwegian University of Science and Technology. The work has been performed throughout my 10'th semester, spring 2006 at Telenor R&D Fornebu. It has been a educational period where I have gained knowledge first of all relative to the thesis performed, but also other interesting cutting-edge technologies elaborated at the R&D department.

I would like to use this opportunity to thank professor Lill Kristiansen for valuable input in the startup and finishing phase of the thesis work. A thank goes also to the people at Telenor R&D and especially the supervisors Knut Kvale and Narada Warakagoda for valuable input and help on implementation. I would also like to thank my fellow MSc. students and lunch mates at Fornebu for sharing good ideas and relevant information.

Fornebu, 16. June 2006

Thormod Schie

Contents

List of Figures

List of Tables

Abbreviations

3GPP	Third Generation Partnership Project
AAA	Authentication, Authorization, Accounting
ASR	Automatic Speech Recognition
EDGE	Enhanced Data Rates for Global/GSM Evolution
GGSN	Gateway GPRS Support Node
GMSC	Gateway Mobile Switching Center
GPRS	General Packet Radio Service
GSM	Global System for Mobile Communications
HLR	Home Location Register
HSS	Home Subscriber Server
HTML	HyperText Markup Language
HTTP	Hyper Text Transport Protocol
ISDN	Integrated Service Digital Network
MMS	Multimedia Messaging System
MSC	Mobile Switching Center
MUST	Multimodal, multilingual information Services for small mobile Terminals
Node B	UMTS Base Station
PCM	Pulse-code modulation
PSAP	Public Safety Answering Point
PSTN	Public Switched Telephone Network

RNC	Radio Network Controller
RTT	Round Trip Time
SGSN	Serving GPRS Support Node
SMS	Short Messaging System
TDMA	Time Division Multiple Access
TTS	Text-to-Speech Synthesis
UE	User Equipment
UMTS	Universal Mobile Telecommunications System
URL	Unified Resource Locater
USIM	UMTS Subscriber Identity Module
VLR	Visitor Location Register
W3C	World Wide Web Consortium
WAP	Wireless Application Protocol
W-CDMA	Wide Code Division Multiple Access

Abstract

This thesis presents a mobile multimodal service platform, which enable users to interact with automated services using a standard mobile terminal in a user friendly and efficient way. The concept of multimodality was introduced to the world of mobile devices and services because one saw the limitations that conventional interaction methods posed. There is a merging trend that people want to be more mobile and have access to different services when on the move. To adapt to the user needs, the service providers try to develop mobile services. The problem is that these services are becoming ever more complex and requires more interaction from the user. The paradox is that the mobile devices these new services are accessible from has not evolved in the same speed. Most of the mobile devices sold on the market today basically comprise of a display and a simple keypad. Thus, to navigate in and operate a mobile service, requires both patience and handiness using a mobile handset.

The multimodal system worked on in this thesis is a speech-centric multimodal platform based on a client-server architecture. The user connect the multimodal client part with the multimodal server part and can thereafter interact with a multimodal service using both speech commands and touch-sensitive display to point at objects etc. in the graphical user interface. As a response to the user queries the system can present the results using both graphics and synthesized speech. This may not sound as a revolutionizing new concept, but what the multimodal interface provides is the possibility to give simultaneous input. I.e. the user may point at a icon on the display while simultaneously input speech commands. These two inputs interpreted one by one will give no meaning, but interpreted together they constitute a reasonable user query. In this way, the user can do interaction based on own preferences. The result of the query is also multimodal, i.e. the system present the results according to the user preference, and the user will hopefully have a better user experience.

The thesis look into multimodality and relevant technology. Further it specifies requirements for the mobile multimodal service. Based on the requirements, a mobile multimodal solution is elaborated. The implementation and the presented solution elaborated in the thesis is based upon a multimodal platform which among others Telenor R&D has contributed to. The original multimodal platform base the communication between client and server on a WLAN connection.

To improve the mobility, functionality for connecting client and server using third generation mobile network technology referred to as 3G, or more specific UMTS, is implemented. Further an analysis on how well the implementation cover the specified requirements is performed.

Finally, considerations about the multimodality and the presented solution is discussed, with emphasize on reliability, usability and openness.

Chapter 1

Introduction

Would it not be nice to communicate with a automated service through your mobile terminal in the same way as you communicate with a human service agent? All the implicit information you communicate with a human agent is totally incomprehensible for a mobile service. When you try to explain or ask the service agent something, you simultaneously point at objects etc. and make certain gestures. For a mobile device to understand and comprehend these gestures may sound as a utopia today. Nevertheless, research are done to make this become a reality, probably not tomorrow, but maybe in a distant future. To get there we need to take one step at a time. The first step on the way to a more user friendly mobile future, the concept of multimodality can be the answer.

1.1 Motivation

Today's information technology devices are generally unimodal, i.e. users are restricted to a single mode of interacting and inputting to them, restricting the value of the new communications channels such as the web. Applications designed to be multimodal allow for much greater freedom of choice when interacting with these mobile devices, be it keystrokes, hand writing or speech commands.

There is a drive today toward more intelligent and user friendly interaction methods on mobile devices. Mobile operators, service and content providers and mobile users want better options for navigating and interacting with different types of services provided over a mobile network. For mobile users when applying e.g. a WAP service, it feels like looking through a keyhole when viewing content. And when interacting with a mobile service it is painstaking to navigate and 'surf' between different sites, when the keypad is the only tool for giving input to the device.

Service and content providers are very restricted when they develop content and services for mobile devices. This is because the mobile terminal and appurtenant technologies set restrictions on how content can be displayed This is reflected in the content they provide. Often are the WAP pages and mobile services out of date that the users instead go to standard web pages. Although this make the 'keyhole factor' even greater, at least the content are more up to date. Whether it because of the poor public interest or the low priority from the providers that causes this vicious circle is unknown. But by introducing a more user friendly platform for providing content and services could improve the popularity of mobile services.

The mobile operators have also incentives for providing a better platform for mobile services. A recent headache for the mobile operators have been to generate revenues from their large investment in the new third generation mobile networks. The problem with the current 3G networks are that they do differentiate themselves compared to the older GSM network. Thus, the customers do not see any good reasons for converting to the 3G network. Based on this the operators need to develop attractive services which exploits the capabilities in the 3G network to attract customers to change. Additionally, if mobile services becomes more attractive among the public, this will generate more traffic in the network, resulting in increased revenue.

Another motivation factor is that new and more user friendly interaction methods can contribute to that certain user groups that currently can not benefit from services that should be publicly available today. People with different disabilities may have different issues that prevents them from using these services. The concept of universal design [13] is very important in this context. The principles of universal design is that products, services and environments should be developed with the aim to accommodate every people in the design phase, instead of doing necessary adjustments to the service or product when it is finished. Universal design include the concept of multimodal interaction, thus concept of multimodality can contribute to make the every day life for many people easier than it is today.

1.2 Scope and Limitations

The thesis will focus on the concept of multimodal interaction and how this concept can be introduced on mobile devices. It will be elaborated a solution for a mobile multimodal service. The scope of the thesis will be toward the multimodal platform and the infrastructure that enables the multimodal functionality. It will also be looked at different issues that can restrict and/or promote the usage of multimodal interfaces. It is out of scope to look at different applications built upon a multimodal platform, however for illustrative purposes example services will be explained. The multimodal platform introduced are quite extensive so to confine the scope of the thesis, elaboration attached to certain parts of the multimodal platform is left out.

Which parts and reasons for it will be explained when it is relevant.

1.3 Thesis outline

The thesis structure is divided in chapters. Chapter 2 introduce the concept of multimodality and describes related work. In chapter 3, technology that is required to enable multimodal interaction is described, but also technology that can improve the usability and performance of a multimodal system is included. Chapter 4 is basically divided in three parts. First a general overview and a detailed description of a multimodal platform is given. Then requirements set to a mobile multimodal service based on different scenarios given, are specified. As a response to the requirements, the last section is devoted to implementation details presenting a solution that will try to fulfill the specification. Chapter 5 presents results based on the specified requirements. Chapter 6 will discuss the presented solution and issues in connection with the solution. It will also discuss possible enhancements. Chapter 7 will conclude the thesis.

Chapter 2

Background

This chapter will go through different background information that is relevant for the thesis. The concept of multimodality is very central in this thesis and will be thorough explained in the first section. The subsequent section will give information about the system worked on in this project.

2.1 Multimodality

H. J. Charwat [8] defines modality as: "Perception via one of the three perception-channels. You can distinguish the three modalities: visual, auditive, and tactile (physiology of senses)."

This section will discuss multimodality in general and reasons why the concept is interesting in advanced telecommunication services. Multimodality is a concept that allows humans to use different modes of interaction according to what is most suitable in different situations. To avoid mixing up the words, throughout this thesis the definition of 'output' is information delivered from machine to human and 'input' is information communicated from the user to the machine. Multimodal must not be confused with multimedia. Media is defined as the representation format for the information or content that is to be conveyed. For example you have visual media such as text and pictures and auditive media such as speech and music.

2.1.1 What is multimodality?

Multimodality means the use of more than one modality. This means that a user should be able to use more than one mode when giving input to a service. In the same manner should the system be able to give output to the user using more than one modality. The definition of multimodality sets no restriction on whether the modalities are applied simultaneously or sequentially. This will be described in more detail in section 2.1.3.

It should be noted that multimodality is restricted to those interactions which constitute more than one modality on either the input or the output side. This means that traditional user interaction with a computer which comprise use of keyboard for input (tactile mode) and monitor (visual mode) for output is not considered as multimodal interaction. The different modalities is defined as communication channels between user and machine.

The three modalities visual, auditive, and tactile is associated with three of the five human senses. Visual corresponds to the eyes, auditive to the ears and tactile to the sense of touching. The two remaining senses are not adaptable for use in telecom services on small handheld devices, at least today. The multimodal research in this project is constrained to speech-centric multimodality with two input modes; speech and touch and two output modes; speech and vision.

Figure 2.1: Illustration of different modalities

2.1.2 Why use multimodality?

Multimodality is a interesting concept with regards to mobile terminals such as PDAs and smartphones. These terminals offer limited screen space and no mouse and/or full-fledged keyboard functionality.

Services and applications developed for mobile terminals provides enhanced and complex graphical user interfaces and require much input from the user. Due to the limited capabilities on a handheld terminal, these services become difficult to use with current interaction methods. With a multimodal interface new methods for interacting with services becomes available.

When humans communicate with each other, we use co-ordinated speech, gestures, body language and facial expressions, and we combine different input senses such as vision and hearing. Communication between humans is by nature multimodal, and it is natural to use different modalities simultaneously. Monomodal communication restrict humans to communicate in their natural way. When for example sending text-messages, typical gestures will not be transferred to the other party and misunderstandings may appear. Different means is created to avoid this problem. E.g. a currently popular feature in instant messaging conversations or SMS messages are smiley's expressing different moods and are sent together with text to show whether the person is e.g. ironic, happy or sad.

Another monomodal service is plain old telephone service. When talking with a service agent of some kind one usually need to use pen and paper to write down information received. If the telephone also had a visual mode, the agent and customer could share this interface to exchange important information or be helpful when explaining something via the visual interface.

There is several areas of application where multimodal interfaces could be very helpful and only the imagination sets the limit. Different users have different needs. The subsequent sections will illustrate the use of multimodal interface for different user groups.

2.1.2.1 Usage for non-disabled persons

There are many challenges today for accessing wireless content and applications. The usage of handheld terminals such as mobile phones and PDAs are increasing and the terminals are added with ever more functionality. Mobile phones that at their beginning were restricted to make phone calls, have now the possibility to send and receive small text messages (SMS), send and receive e-mails, browse web pages (WAP) and have instant messaging dialogs. All these services require output to the user and input to the terminal. But the keypad is a poor input mechanism for typing, even though there have been developed different techniques such as predictive text technologies; T9 or iTap to simplify the writing. To keep the mobile phone at a reasonable size the display is kept quite small which reduce the user friendliness of viewing text, pictures and video. The exploding amount of content and information which are made available for mobile terminals makes it difficult to navigate through and get a clear overview of the information browsed. In this way multimodal technology can improve the user interfaces for input and output of data content to provide a faster, easier and more interactive experience for mobile users. Multimodal interfaces allow users to select the most appropriate and suitable input and output modalities for interacting according to user context and enhance the user friendliness of telecom services.

2.1.2.2 Usage for disabled persons

There are quite different issues that appear for persons with different types of disabilities. Depending on the disability, multimodal interfaces can help disabled person using teleservices. The authors of [18] did some research revolving this topic and they found that that map-based information service on a mobile terminal was proved to be useful for severe dyslectic [1] and an aphasic [2] persons. A quotation from the conclusion of [18] reflects their discoveries:

> The severe dyslectic and aphasic could neither use the public service by speaking and taking notes in the telephone-based service nor by writing names in the text-based web service. But they could easily point at a map while uttering simple commands. Thus, the multimodal interface is the only alternative for these users to get web information.

It is not fair to put this complex group of people in to one segment, but the important message here is that people with different handicaps need adjusted equipment fitted to their requirements. The concept of multimodal interface implemented into telecom services can help many of these people to become more self-reliant.

2.1.3 Combining Multimodalities

Different input and output modalities can be combined in several different ways to create a multimodal interface. The World Wide Web Consortium (W3C) has defined three different ways of combining multimodal inputs and outputs: Sequential, uncoordinated simultaneous and coordinated simultaneous multimodal input/output. It is important to distinguish the three scenarios because the complexity of each implementation is very different. In the following sections the different ways are explained more thorough.

2.1.3.1 Sequential Multimodal Input/Output

This is the simplest form of multimodal interaction, where input and output from different modalities are interpreted separately. In sequential operation the input modes cannot be used

[1] Having impaired ability to comprehend written words usually associated with a neurologic disorder

[2] Partial or total loss of the ability to articulate ideas or comprehend spoken or written language, resulting from damage to the brain caused by injury or disease.

simultaneously. Figure 2.2 illustrate that each interaction must be done stepwise and a explicit switch is required when changing from one mode to another. This means that at any given moment only a single, designated input mode is active. But in the whole interaction more than one input/output mode may be used.

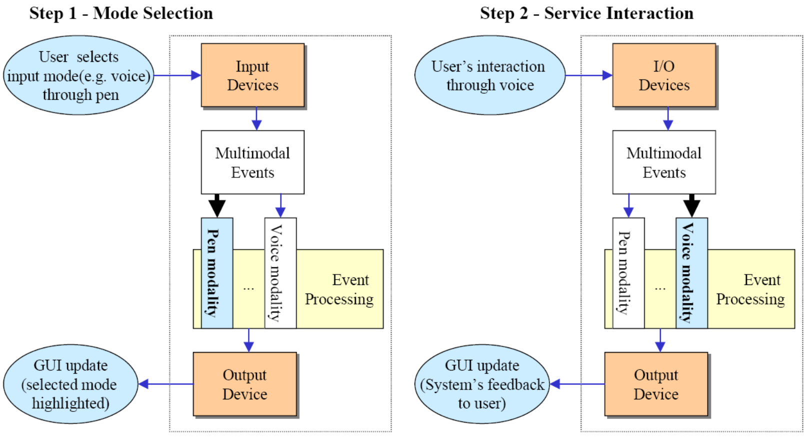


Figure 2.2: Illustration of sequential multimodal interaction [7]

2.1.3.2 Uncoordinated Simultaneous Multimodal Input/Output

This form of interaction is more complex than the sequential mode. In this situation several parallel input modes are active at the same time. This means that the users can choose the input mode they prefer at each dialog stage. But in each turn only one is selected for processing. Which mode is used at each turn can be decided according to different criteria, such as the first mode to start or that one mode has priority over the other. Figure 2.3 illustrate the interaction form.

2.1.3.3 Coordinated Simultaneous Multimodal Input/Output

This is the most advanced form of interaction. Also here more than one input mode is available simultaneously, but in contrast to the uncoordinated simultaneous mode, here all inputs from the different modalities are collected within a time window and interpreted. As figure 2.4 shows, the resemblance with the uncoordinated simultaneous mode form is striking. But the difference is how the modalities are collected and interpreted. In the uncoordinated simultaneous mode the events are interpreted one by one. In the coordinated simultaneous mode the events are combined to create a query to the multimodal system. The fusion process, which is described in the next section, elaborate the partial information coming from the different channels and create a result.

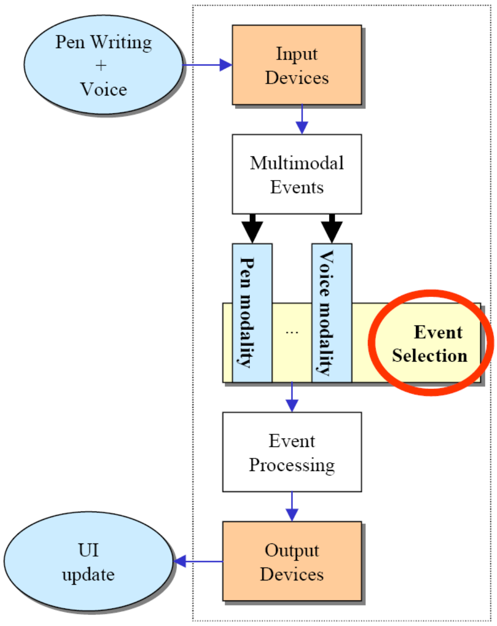


Figure 2.3: Illustration of uncoordinated simultaneous multimodal interaction [7]

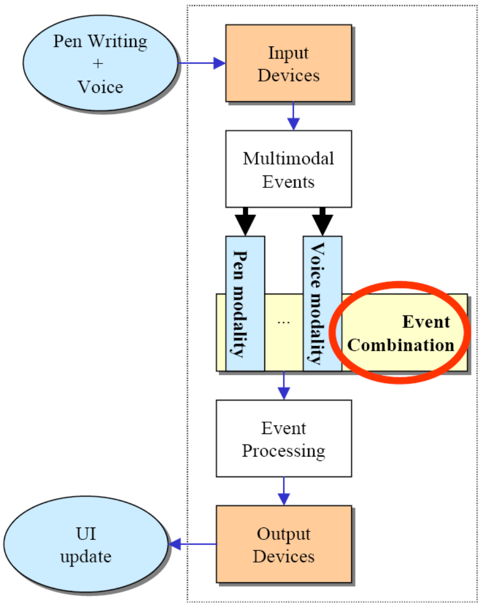


Figure 2.4: Illustration of coordinated simultaneous multimodal interaction [7]

2.1.4 Fusion and Fission

Fusion and fission are important concepts in the context of multimodal interaction. As figure 2.5 shows, the fusion process is done at the input side of a multimodal system and can be explained short for several input channels to a single semantic stream. Fusion is defined in reference [7] and the reference gives a good explanation of the phenomenon:

> In the context of multimodal interaction fusion refers to the combination of several chunks of information, to form new chunks, possibly of a higher order. The information chunks to be fused may or may not originate from distinct input channels or from distinct contexts. For example, the sequence of events 'mouse-down, mouse-up' that occurs in the palette of a graphics editor comprises two information chunks that originate from the same input channel and from the same context (i.e., the palette agent). They are combined within the context of the palette agent to form a higher information chunk (i.e., the selection of a geometric class). Pointing with a pen at a file object on the screen and the spoken word 'open' are an example of information chunks originating from different input channels. After fusion the interpretation is 'open file'.

The fission process is not as advanced as the fusion process. The process is done at the output side and is the opposite process of the fusion, i.e. a single semantic stream is split into several output channels representing different modalities. From the same reference as for fusion, fission is defined as:

> In the context of multimodal interaction fission refers to a kind of decomposition of complex chunks of information into multiple chunks that may or may not appear to be simpler. The resulting chunks may be presented through different output channels (sequentially or simultaneously) or to a single channel (in which case sequential presentation is the only option). The 'decomposition' of the information may result in redundant representations. For example, the information chunk 'Tour Eiffel' can be rendered on a screen by displaying a picture of the structure accompanied by a text label. The picture is rendered through the graphical channel, the textual label through the text channel. The name may also be spoken, adding a third channel with essentially the same information.

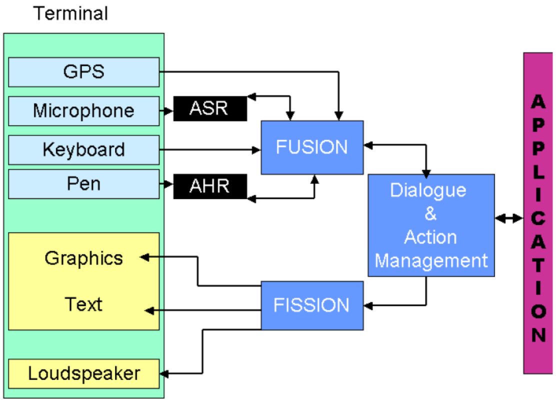


Figure 2.5: Illustration of the fusion and fission processes in a multimodal system [7]

2.2 The MUST demonstrator

The system and platform used in this assignment has been involved in different research projects. It started out as a project managed by EURESCOM [3].

The project was called "MUST - Multimodal, multilingual information Services for small mobile Terminals" [11]. The project was a cooperation between Telenor R&D, France Telecom and Portugal Telecom. The aim of the project was to do research revolving the concept of multimodal services on mobile terminals.

During the project a demonstrator was developed. The main purpose of the demonstrator was to implement an experimental multimodal service in order to study and evaluate the interaction with and the appreciation of such a service by real 'naive' users, hereafter called 'Must demonstrator' or just 'MUST'. A framework for the backend server system and the client part for providing a multimodal service was developed. Initially, the demonstrator provided a tourist guide service for the city of Paris.

The framework for a multimodal service developed in the EURESCOM project was continued in the Norwegian 'Brage Project' [16] supported by the The Research Council of Norway. This project is still running[4] and is a cooperation between Telenor R&D, SINTEF ICT and the Norwegian University of Technology and Science (NTNU). The project's main focus is speech technology and dialog-based human-machine communication.

[3] EURESCOM, the European Institute for Research and Strategic Studies in Telecommunications, performs collaborative R&D in telecommunications. The EURESCOM shareholders are mainly the major European network operators and service providers.

[4] As of June 2006

2.2.1 Different Applications

The ‘Must demonstrator’ has undergone different adjustments, changes in functionality and user interface during its lifetime. But the underlying multimodal functionality has been the same.

MUST is basically a framework, which different multimodal applications can be built upon. Several of the components in MUST can be reused to create new multimodal applications. As mentioned two different applications has been tried out using the multimodal architecture, the EURESCOM and Brage implementation, which will be described in the subsequent sections. Note, this chapter will not go through details about the underlying architecture and technology running the system. This will be presented in chapter 4.

2.2.1.1 Paris Tourist Guide

During the EURESCOM project, the system provided a tourist guide service for the city of Paris. More information about the tourist guide service can be found in reference [10]. The idea was that for tourists coming to Paris, it could be difficult to orientate themselves. There are so many different tourist attractions in Paris, and to be able to locate and get information about them, some help is needed. Of course, a map is a sufficient measure. But to present the city and it’s attractions in a more interesting way, new technology needs to be applied.

A tourist is then equipped with a PDA connected to the MUST server. The tourist log in to the service and is presented with a high level map of Paris, showing some of Paris’ attractions. The user may then make use of either pen, speech or both to navigate through the application. The user can click on the icon for the Eiffel Tower and ask simultaneously ‘What is this?’ The system will then combine the two inputs, interpret it and generate an answer based on available information. At last the system will present the information retrieved about the Eiffel Tower for the user comprising graphics, voice or both.

2.2.1.2 Bus Demo

The service implemented in the Brage project differ from the EURESCOM project. In the Brage project the service provided is a public bus information system, also called ‘Bus demo’. The service is based on a Norwegian public transport information service for the area in and around Oslo called ’Trafikanten’ [35] The service is available on both web and WAP, in addition to public information desks and a telephone service.

What the application provide is a new mobile multimodal user interface to the existing web-service. When the user log in to the service, a map overview of Oslo is presented. The user will then have different possibilities for using the application. For a detailed explanation of the service, see appendix B.3.

2.2.2 Operating Methods

The multimodal system worked on in this project is a speech-centric system, i.e. it is speech combined with input using a stylus pen on a touch-sensitive display that constitute the interaction methods for navigation through the service. To describe the different options for user interaction with a multimodal application, three different scenarios are illustrated in subsequent section with basic in the Bus Demo service. The user has the possibility to use a interaction method that suits the environment and user preference for any given moment.

2.2.2.1 Pure speech oriented

When the user logs in to the service, the user is prompted with a welcoming message. The user may then directly ask the service: “When does the next bus goes from Fornebu to Majorstuen?” The terminal will continuously transfer all sound to the multimodal server, but a voice activity detector will sort out speech from general background noise. Next the voice query is interpreted using ASR to deduce what the user asks. Based on this interpretation an answer is created in response based on available information. The information presented to the user comprise of graphics and text on the display which is at the same time read aloud using TTS, hence it is multimodal output to the user.

2.2.2.2 Pure pen oriented

It is also possible to interact with the service using only the pen as input. In this scenario the user zoom into a map-segment and clicks on the bus station he wants to leave from or go to. Next the user needs to find the station in which he wants to travel to or from. Whether it is a ’To’ or ’From’ bus station can be indicated by pressing a ’To’ or from ’From’ button available in the user interface. The user is then prompted with a question whether he wants to get information about the next bus or not. When confirming to this question by pressing ’Yes’ on a combo-box, the input is elaborated and an answer is presented to the user in the same manner as for pure speech.

2.2.2.3 Combined speech and pen

The last option is a combined speech and pen mode. This scenario is more advanced and it really utilize the concept of coordinated simultaneous multimodal input as explained in section 2.1.3.3. Using this approach to interact with the service can put the user up to some challenges. Normally, when a human interact with a machine of some kind it is done by only one modality. Therefore it may be awkward to use both speech and pen, especially simultaneous. Nevertheless, in this scenario the user has several options for giving input to the server. The user may now click into the map-segment where he wants to leave from or go to and do one of the following:

- The user can say: "When does the next bus go from 'here' to Majorstuen?" 'Here' then refers to the bus-station icon the user clicks while saying the sentence.
- Or the user can say "When does the next bus go from Fornebu to 'here' ?" Then 'here' is the same as the latter example except that now it is referred to as the destination address.

In this scenario the result is also presented with text and speech, hence a multimodal output. The difference from the other user scenarios are that in this case the fusion process is different. That is because this interaction method is based on truly coordinated simultaneous multimodal input. The fusion process is fed with both speech and pen input and needs to resolve an answer based on the two inputs. It is in this scenario where the multimodal system is put up to the test to show its potential as a multimodal service.

2.3 Related work

There is at the time much research devoted to the area of multimodal interaction. As indicated earlier, much of the standardization work in the area of multimodal interaction is done within the W3C organization. W3C elaborate and recommend standards to be followed by both research departments and commercial companies.

Open Mobile Alliance (OMA) is a joint industry forum comprising both mobile systems manufacturers, mobile operators and software vendors. The aim of OMA is to ensure service interoperability across devices, countries, operators and networks. One example of such a interoperable service is the SMS application or simply standard voice calls.

Within this forum requirements regarding multimodal service has been elaborated. The scope of the document 'Multimodal and Multi-device Services Requirements' [24] is to specify requirements for multimodal services that manufacturers and operators should follow. This will help to ensure interoperability across devices and networks. History has told us that to achieve successful applications it is important to elaborate open standards that all manufacturers must follow. Using this strategy, the competition between manufacturers are focused on creating good products and services within the limits of the standards.

In addition to the research work done in projects as mentioned earlier i.e. MUST project and Brage project, other research projects initiated by different commercial participants exist. The companies Motorola, Opera Software ASA and IBM have worked out a proposal for a multimodal markup language standard called XHTML+VoiceXML (X+V) [15]. This proposal provides a way to create multimodal web applications, i.e web applications that offer both a visual and a voice interface. Their work is still on the drawing board, but some prototypes and demos have been developed to illustrate the concept.

The concept of multimodal interfaces have definitively reach beyond the research stadium. Different companies worldwide have released commercial versions of multimodal platforms. Among them are Nuance, which is a software company specializing in speech technologies. Nuance has developed a multimodal platform called Xmode [22], which has a client-server architecture, where the client is installed on a handheld device. The system support different operating systems and different types of handheld devices. The communication between client and server can be transferred over different packet-based mobile networks.

The company Kirusa [17] has devoted their effort solely to the concept of multimodality. Kirusa develop and sell multimodal wireless platforms that mobile operators and service providers can apply to offer multimodal applications to their customers. Their solution supports all major mobile networks.

Chapter 3

Relevant Technology

This chapter will go through different technologies relevant for this project. Some of the technologies described has directly relevance with the project, but other a more general and is included for illustrative purposes.

3.1 Mobile wireless networks

There exist a wide variety of cellular technologies around the world, due to local regulations and history. The author have constrained this section to look at some of the most popular cellular technologies. These standards can be defined under the collective term 'GSM-family', which comprise the standards GSM, GPRS, EDGE and UMTS. The following section will describe some different mobile network standards available in the market today. A more detailed description where the characteristics of the different network are compared is specified in the following subsequent sections.

3.1.1 GSM

GSM is called a second generation mobile systems (2G). It was one of the first digital mobile network developed and thus has become a widely adopted standard across the world. The GSM standard uses TDMA, which is a narrow band solution. The radio-interface is divided into frequency channels and each channel is divided into time slots providing eight channels per radio frequency channel.

GSM was originally developed to provide circuit-switched voice and data connections . The GSM standard has been continuously developed, and new enhancements has been added. Among them improvements for providing better data-rate capabilities. Original, the GSM technology provided poor data-rates. It was apparent that this had to be improved and new modulation techniques and the possibility to use several time slots simultaneously provided better data-rates. Nevertheless, the utilization of the radio resources is poor, due to the circuit-switched technology which reserves the full bandwidth during the lifetime of the connection.

3.1.2 GPRS

General Packet Radio Service (GPRS) was developed to provide packet-switched data service to the existing GSM network. GPRS is more efficient for data transmission than the circuit-switched GSM. This is because GPRS utilize the network capacity better by dynamically share the available bandwidth between multiple users. GPRS is given a lower priority than circuit-switched services in GSM network, thus in busy areas/cells GPRS may experience poor bit rate. Most GSM terminals sold today, support GPRS. Even though GPRS provides data capabilities, the services available suffer from low bit rate. This is especially true when using mobile Internet-services such as web-browsing and streaming multimedia. GPRS does not support real-time services, thus real-time services such be handled by GSM CS. More on this issue in section 3.2.1

3.1.3 EDGE

Enhanced Data rates for GSM/Global Evolution, also called EGPRS (Enhanced GPRS) is an enhancement to GPRS, thus a part of the GSM standard, providing better data rates. EDGE is basically an upgrade of the GPRS service, i.e. an upgrade to the air interface between the terminal and the network. EDGE can provide three times higher data rates than GPRS, using a more effective modulation and coding scheme. EDGE is currently being deployed on a high scale at the moment and finally gives mobile users a satisfactory transmission capacity for advanced telecom services and Internet-applications.

3.1.4 UMTS

Universal Mobile Telecommunications System (UMTS) is one of five candidates specified by IMT-2000 (International Mobile Telecommunications-2000 as a global standard for third generation (3G) wireless communication, defined by the International Telecommunication Union (ITU). UMTS is standardized by the Third Generation Partnership Project (3GPP) [27]. UMTS is a multi service network with the possibility to provide TV, video, high-speed multimedia data services and mobile Internet access.

UMTS was developed with the aim to succeed GSM. As it will be explained later, GSM and UMTS use much of the same components. This is done to enable easy hand-over between the two networks in areas where one of the networks suffer from poor coverage. UMTS differs from GSM in that it employ a whole different radio interface called W-CDMA. This is very different from GSM's TDMA, in which it offers new features and better capabilities. UMTS is an immature network technology and continuous change is necessary. UMTS has been standardized by 3GPP in different releases. The first version called Release 99 is the version that is most deployed today, but the standard is ever evolving and new releases are set for the future. Release 4, 5 and 6 has frozen, and is ready to deployed in new UMTS network equipment. These new release will provide better uplink and downlink data rates, voice over IP (VoIP) services and better support for multimedia services over the mobile network [27].

3.1.5 A joint GSM and UMTS Network

This section will describe the topology and characteristics of the networks belonging to the GSM-family. The network developers and manufacturers have made sure that much of the GSM equipments that operators have invested a great deal of money in could be reused when deploying the UMTS network. This have resulted in that the current versions of the networks are congruent and operators can take advantage of this by reusing many of the existing components in their GSM/GPRS/EDGE network when deploying UMTS.

On this behalf the technical details about the mentioned networks will be explained in the same context. The joint network architecture is called 3GSM. 3GSM is the latest addition to the GSM family. The technology on which 3GSM services are delivered is based on a GSM system enhanced with a W-CDMA air interface [38].

Figure 3.1 shows a combined GSM and UMTS Release 99 (UMTS99) network architecture. The network provide services for the whole GSM family.

The mobile network consist of three logical interacting domains; Core Network (CN), Radio Access Network (RAN) and User Equipment (UE). As the figure shows, the CN is common for the both networks, the RAN is different for the two standards GSM and UMTS and mobile equipment can be common for both. The following sections will go through the specified domains.

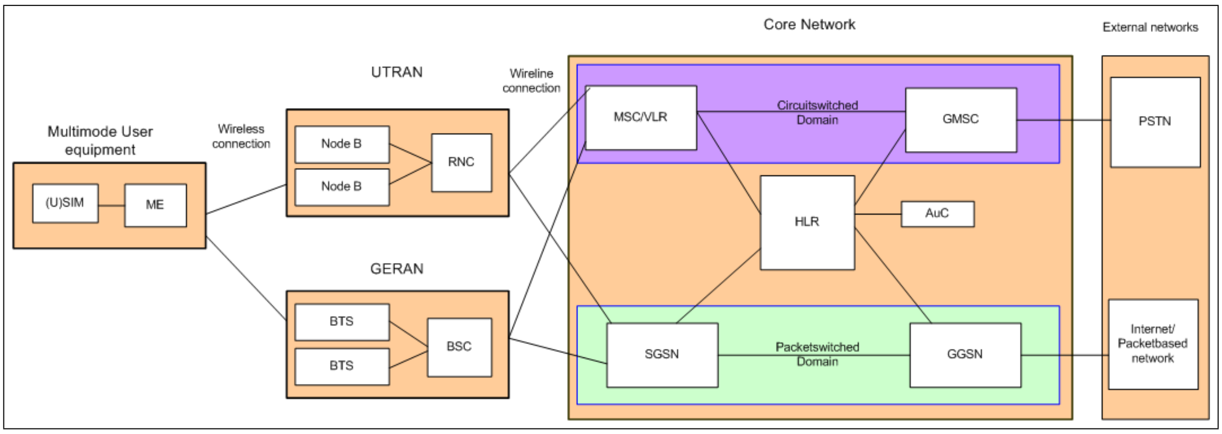


Figure 3.1: A logical GSM and UMTS network topology

3.1.5.1 Core network

The core network (CN) for UMTS99 is more or less the same as for the GSM core network, though with some upgrades for supporting UMTS. The following nodes are reused in a common UMTS and GSM CN:

- MSC Mobile switching center
- AuC Authentication center
- HLR Home location register
- VLR Visitor location register
- SGSN Serving GPRS Support Node
- GGSN Gateway GPRS Support Node

The main function of the CN is to provide switching, routing and transit for user traffic. This is made possible because UMTS99 and GSM are using the same signaling system and the same protocols for transmission of user data. Core network also contains the databases and network management functions. As figure 3.1 shows the core network is divided into two domains, a circuit-switched domain and a packet-switched domain.

The circuit-switched (CS) domain is known to provide services such as voice calls. CS has been the dominant technology for several years. The technology is well known and have been preferred when implementing real-time applications such as voice calls. This is because it preserve resources before the connection is established, thus the connection is guaranteed a minimum of resources. The telecom industry has used the experience acquired from developing and operating wire line services such as PSTN and ISDN when developing the GSM system, hence many of the standards are continued in the GSM specification.

The packet-switched (PS) domain is known to provide services such as IP-based traffic known from the world-wide Internet. With the increasing popularity of Internet-based services such as WWW and email, the traffic pattern have changed. This has made the CS technology less attractive for packet-based traffic, because capacity are left unused in periods when the user is passive.

For telephony over wireless networks, CS is still the preferred technology because it is more effective than packet-based technology. This is because of the overhead that comes with packet-switching technology. Telephony transferred over a packet-based network generates small packets to deal with the real-time aspect of telephony. Hence, it is generated many packets each with a header resulting in much overhead data.

3.1.5.2 Radio Access Network

The Radio Access Network (RAN) part is where UMTS and GSM differs the most. Based on the experience gained from deploying and operating the GSM network and due to the different radio interface, it was made some changes to the UMTS RAN compared to the GSM network. As figure 3.1 shows, the corresponding modules in the UTRAN has different names and functionality compared to GERAN.

- BTS - Base Transceiver Station is a two-way radio module with an antenna providing connectivity for mobile phones.
- BSC - Base Station Controller provides the intelligence behind several BTS' and handles allocation of radio channels, control handover between different BTS etc.
- Node-B corresponds to the GSM's BTS, but have a new name due to using a another radio interface.
- RNC - Radio Network Controller corresponds to GSM's BSC but is more complex and provides more functionality compared to the BSC.

The original GSM Radio Access Network was built to support speech calls and slow data transmission using circuit-switching. The internal links in the network had a maximum bandwidth of 64 kbit/s. When implementing GPRS, the need for higher capacity became apparent, and upgrades to the network were added.

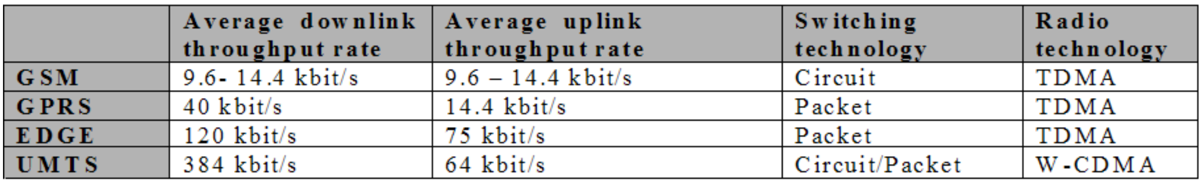
UTRAN was build from scratch with the aim to support both PS and CS technology. Using W-CDMA on a higher frequency than GSM, the cell-size of UMTS is smaller. This results in more handover between cells. Due to this, more of the handover functionality was moved outwards in the network providing better capacity and scalability.

3.1.5.3 User Equipment

Almost all UMTS-supported user equipment on the market today is so-called dual-mode UMTS and GSM compatible terminals. This means that a user will be handed over to GSM in areas which lacks UMTS coverage. Nevertheless, the GSM network will not be able to provide the same range of services as UMTS. More details on the user equipment part will come in later sections.

3.1.6 Characteristics and Capabilities in GSM and UMTS

As mentioned earlier, the different networks are closely related. Basically, they utilize the same network and the biggest difference is the type of radio-interface they use. This knowledge is important when discussing the different standards. GSM with it's upgrades GPRS and EDGE, and UMTS all provide more or less the same services to the customers, but a few features differentiate them. It should be noted that while GSM provides circuit-switched services and GPRS/EDGE provides packet-switched services for GSM, UMTS is a collective term for both circuit- and packet-switched services. Table 3.1 shows some basic differences between the standards.



	Average downlink throughput rate	Average uplink throughput rate	Switching technology	Radio technology
GSM	9.6- 14.4 kbit/s	9.6 – 14.4 kbit/s	Circuit	TDMA
GPRS	40 kbit/s	14.4 kbit/s	Packet	TDMA
EDGE	120 kbit/s	75 kbit/s	Packet	TDMA
UMTS	384 kbit/s	64 kbit/s	Circuit/Packet	W-CDMA

Table 3.1: Characteristics of different cellular technologies [38].

3.1.6.1 Handover

Handover is a mechanism that ensures always connectivity for a mobile users. When a user is connected to a base station and he moves away from that specific base station, the signal will get weaker, and at some point, the user needs to connect to another base station with better signal strength. This action is critical when the user is connected via a CS or a PS connection, because at a specific instance, the traffic needs to be routed a different path through the network. The handover mechanism is usual within a network, but when they started to develop the UMTS network, they saw that it was necessary to develop handover mechanisms also between different mobile networks. To ensure full connectivity for 3G customers, a mechanism to hand 3G customers over to a GSM network in areas where the 3G network still was not fully deployed was develop.

Handover needs to be performed when the user is connected to the CS domain or the PS domain. But the characteristics of these two connections are different, thus the requirement set to the handover in the two domains are different. Telephony are currently handled over a CS connection. For instance during a telephone call, the delay the user experience when being handed over within a mobile network or between e.g. a UMTS and GSM network should as short as possible. Preferably without the user noticing it at all. For a PS connection, which usually is used for non real-time data transfers, the requirements set to the handover is less strict than for CS connections.

Figure 3.2 shows how an example of a PS domain where the user can be covered by either UMTS or GPRS coverage or both. Typical in areas which is covered by both GPRS and UMTS data, the subscriber will receive UMTS coverage because this is the preferred network. When the subscriber moves out of an area of UMTS coverage, the network and terminal will perform a seamless handover to the GPRS network [1].

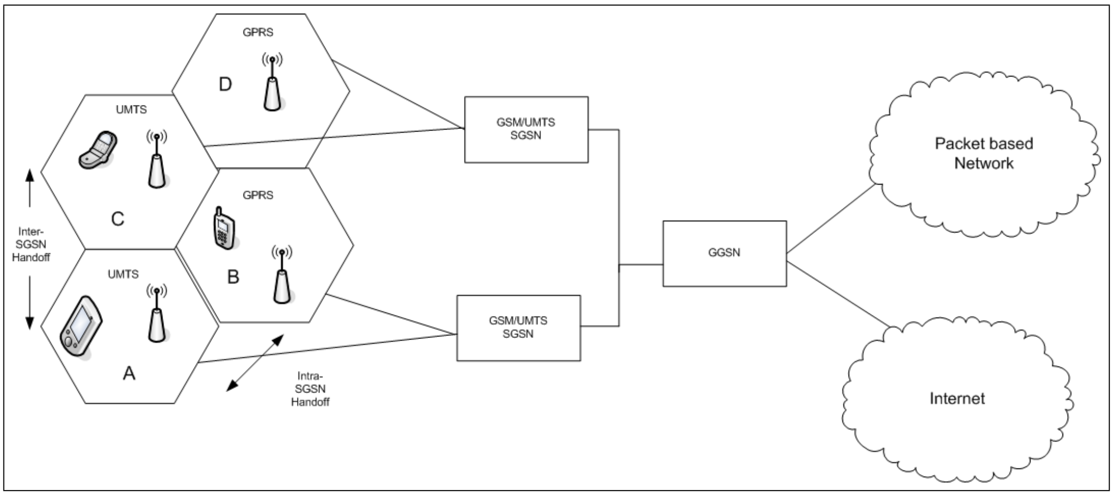


Figure 3.2: Overview of a combined UMTS/GSM packet-switched domain

When the user moves from cell A to cell B, the terminal is handed over from one base station to another, but the SGSN is serving both cell, so the terminal experience a so-called ‘Intra-SGSN handoff’, meaning it is only changing base-station. When a user is moving from cell A to C, the

[1] Assuming there is GPRS coverage available.

terminal must change SGSN, because the new cell is served by another SGSN. This is called a ‘Inter-SGSN handover’. This is obviously a more complex procedure than the intra-SGSN handover.

As figure 3.2 shows, the two SGSN is served by a common GGSN. The GGSN functions as a gateway between the two SGSNs and a specified packet-based network such as the Internet.

As indicated in the CS domain section, more and more traffic is handled in the PS domain. Due to rapid deployment of IP and IP based services, wireless carriers consider the PS domain as a area with biggest potential for generating revenues in the future. The development in the PS area is also good news for the customers who are getting ever more dependent on IP-based services.

3.1.6.2 Throughput

As table 3.1 shows the different standards provide different transmission rates. Because the circuit-switched technology is not very suitable for data transmission the GSM part will not be discussed further here.

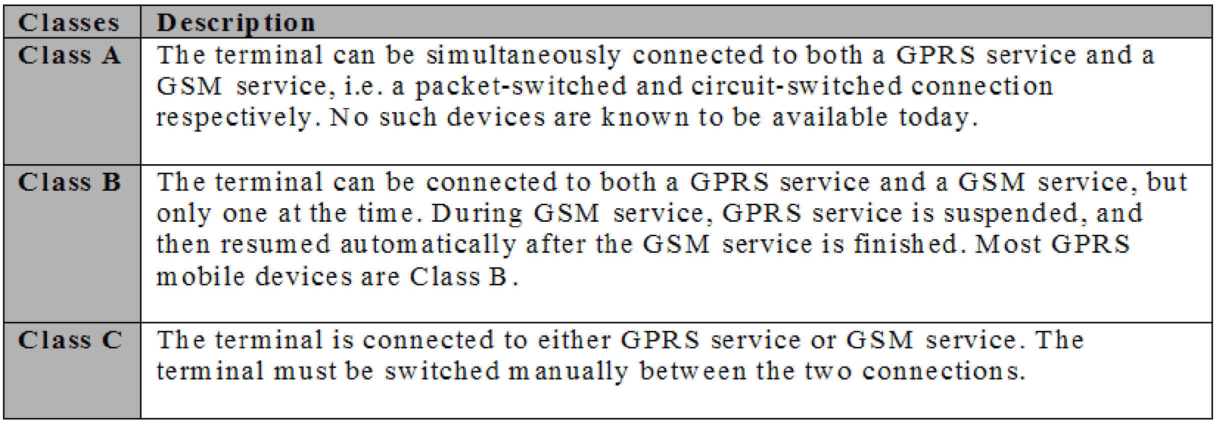
The throughput for GPRS, EDGE and UMTS is generally asymmetric. The resources available is limited, so the downlink direction has been prioritized. This follows of the typical traffic pattern for a regular user where applications such as WWW, email and multimedia-services require mostly downlink transmission.

The quoted up- and downlink rates in the table only serves as an indicator on what the different standards can offer. The available rates depends on many different factors such as type of terminal, infrastructure, network manufacturer, traffic in network, distance from base station etc.

3.1.6.3 Simultaneous circuit- and packet-switched connections

As of today, very few services use simultaneous real-time flows over circuit-switched connections and interactive flows over packet-switched connections. This is because no GSM terminals have supported this feature. As table 3.2 shows, the feature is specified in the standard, in term of GPRS class A. But due to radio interface technicalities it would require two radio transceivers in the mobile device to support this feature. This is uneconomical and the market has not had a high demand for it.

In recent years new promising methods has been discovered which could enable simultaneous circuit- and packet-switched connections for GSM networks. The technology is called Dual

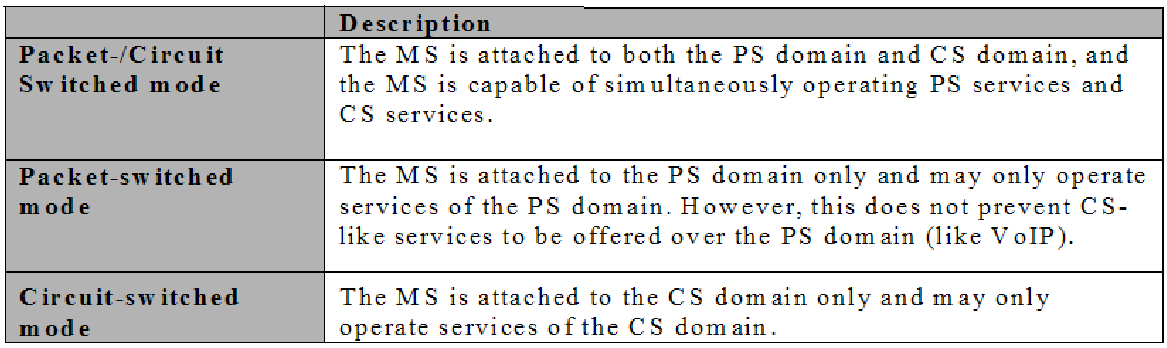


Classes	Description
Class A	The terminal can be simultaneously connected to both a GPRS service and a GSM service, i.e. a packet-switched and circuit-switched connection respectively. No such devices are known to be available today.
Class B	The terminal can be connected to both a GPRS service and a GSM service, but only one at the time. During GSM service, GPRS service is suspended, and then resumed automatically after the GSM service is finished. Most GPRS mobile devices are Class B.
Class C	The terminal is connected to either GPRS service or GSM service. The terminal must be switched manually between the two connections.

Table 3.2: GPRS classes

Transfer Mode (DTM) and much research has been put into the topic [23, 25]. The method does not require two radio-transceivers, making it more rational and cost-effective.

In UMTS, due to the radio interface used, it is fairly easy to implement support for multiple, parallel bearers over the air interface. This enables simultaneous circuit- and packet-switched connections. UMTS equipment is able to work in different modes of operations, see table 3.3.



	Description
Packet-/Circuit Switched mode	The MS is attached to both the PS domain and CS domain, and the MS is capable of simultaneously operating PS services and CS services.
Packet-switched mode	The MS is attached to the PS domain only and may only operate services of the PS domain. However, this does not prevent CS-like services to be offered over the PS domain (like VoIP).
Circuit-switched mode	The MS is attached to the CS domain only and may only operate services of the CS domain.

Table 3.3: UMTS modes of operation

The advantage of having a simultaneous circuit- and packet-switched connections might not be clear. One reason for this is that this feature have been unavailable until the UMTS network was opened. Just recently it was released a service which require both switching technologies simultaneous. The service is a automatic ‘Who is it?’ service delivered by a Norwegian ‘White Pages’ service called ‘Opplysningen 1881’ [1]. The application is installed on your 3G mobile. When you receive an incoming call and the number of the caller is not stored in your of phone book, the service looks up the name in a telephone subscriber registry using a UMTS packet switched data connection and presents it to you in the terminal's display.

This simple application is just one example of a service that exploits this feature, and it will most certainly be released other similiar or more complex services in the future based on this capability. One possible service is during a conversation you want to show to the other party a picture or similar. With the mentioned feature, you can transfer the picture over the packet-switched service, while still talking with the person. More on this feature can be read in the article [23].

3.2 QoS

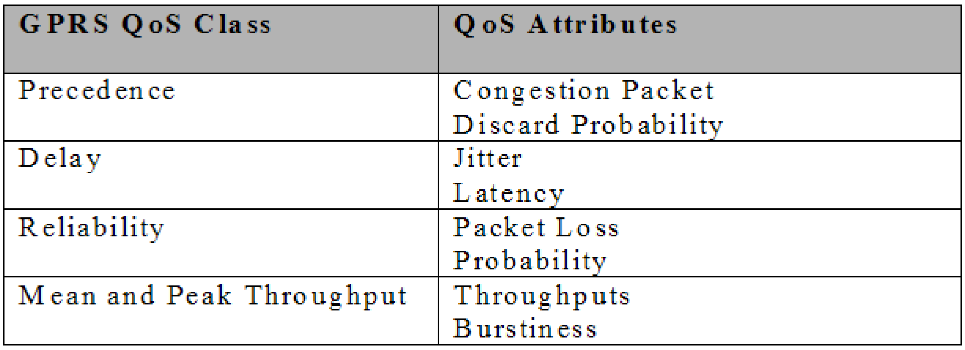
The first packet-oriented network such as the Internet was merely meant for transporting data packets. The only requirement was that the packets reached the correct receiver and that they were free from error. The network was often referred to as a 'best-effort' network. With the ever increasing number of applications developed for the Internet, especially real-time applications, it is necessary that the network can give certain guarantees for the traffic handled. It is not necessary that all types of traffic is getting the same treatment, i.e. different types of traffic may have different requirements. It is this context the concept Quality of Service (QoS) is relevant.

QoS is most relevant in connection with packet-switched networks. For circuit-switched networks, there is established a dedicated channel between two parties, hence such connections are well protected and can provide a stable communication channel. Packet-switched networks on the other hand, can experience different problems as the list below describes. These problems should be handled by certain QoS mechanisms.

- Dropped packets - packets may be dropped en route when traveling through a network.
- Delay - packets can experience delay because it gets held up in long queues, slow links etc.
- Jitter - packets can take different routes between two end-nodes resulting in different delays between each received packet.
- Out-of-order delivery - because packets can take different routes they may appear in a different order to the one with which they were sent.
- Error - packets can be corrupted during the transmission resulting in error in the packet.

3.2.1 QoS in mobile networks

The GSM and UMTS standard have different approaches for implementing QoS in the standards. GSM was originally a pure circuit-switched oriented network. The introduction of packet-switched data services i.e. GPRS and EDGE, specific QoS requirements was defined, see table 3.4.



GPRS QoS Class	QoS Attributes
Precedence	Congestion Packet Discard Probability
Delay	Jitter Latency
Reliability	Packet Loss Probability
Mean and Peak Throughput	Throughputs Burstiness

Table 3.4: QoS properties for GPRS [39]

When elaborating the UMTS standard, it was specified that the network should provided both circuit-switched and packet-switched services. This resulted in extensive focus on providing QoS [6]. Figure 3.3 shows the QoS architecture for UMTS.

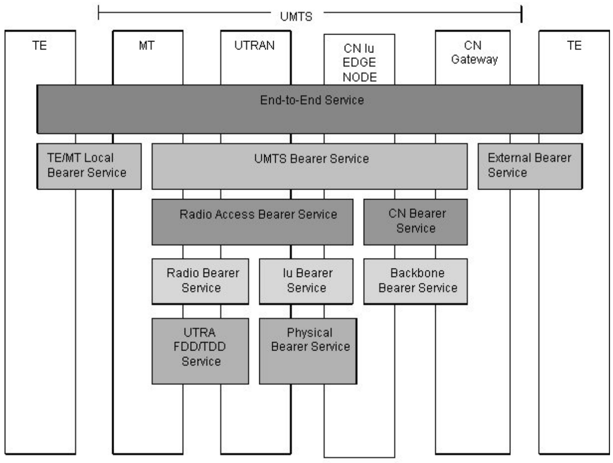
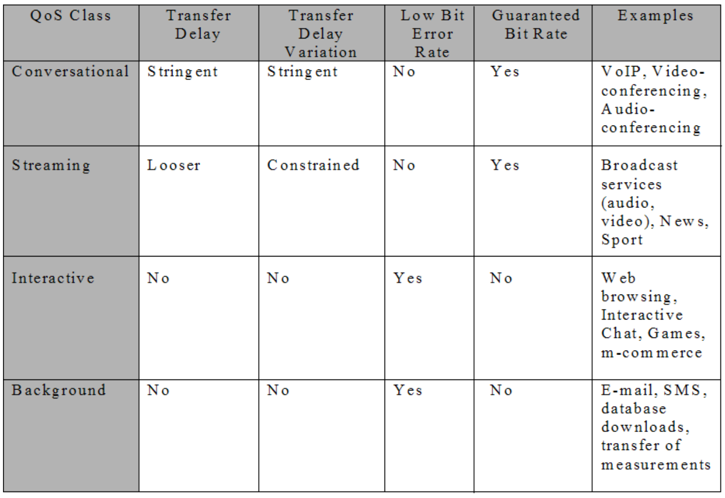


Figure 3.3: UMTS QoS Architecture [26]

UMTS employs four traffic classes for data, see table 3.5.

The different traffic classes have been given different priorities depending on the type of traffic. The classes streaming and conversational is vulnerable to delay. Thus applications belonging to this class are prioritized over less delay sensitive applications such as web-browsing and file transfer.



QoS Class	Transfer Delay	Transfer Delay Variation	Low Bit Error Rate	Guaranteed Bit Rate	Examples
Conversational	Stringent	Stringent	No	Yes	VoIP, Video-conferencing, Audio-conferencing
Streaming	Looser	Constrained	No	Yes	Broadcast services (audio, video), News, Sport
Interactive	No	No	Yes	No	Web browsing, Interactive Chat, Games, m-commerce
Background	No	No	Yes	No	E-mail, SMS, database downloads, transfer of measurements

Table 3.5: UMTS Traffic classes[26]

The QoS concept embraces many areas and challenges. The paper will therefore restrict itself to cover the area of delay in mobile wireless networks.

3.2.2 Delay

Delay is a critical factor in many mobile applications especially for real-time services. The following references deal with delay in mobile networks [12, 28]. Delay is a complex subject and is a result of different parameters such as:

3.2.2.1 Packet loss

Packet loss may result in retransmission of packets, hence the additional delay is added. However, it should be noted that for real-time data traffic which is very sensitive to delay, it is not necessary to retransmit each packet, because a random lost packet now and then can be compensated by the receiver and be unnoticeable for humans. For other traffic types such as file transfer etc. it is necessary that all packets are correctly transfered or else the file will be corrupt.

3.2.2.2 Propagation delay

Propagation delay is the time the signal uses to traverse the media. This delay is dependent on the media in traverse, whether it is through the air or through a wire line such as copper or optical.

3.2.2.3 Processing and routing

As described in 3.1.5.1 packets going to or from a mobile terminal needs to traverse several nodes in the mobile network before it reaches its destination. Figure 3.4 shows the different nodes that the data packets must traverse on its way through the packet switched domain. If the receiver is also terminal connected to mobile wireless network, the packets will need to traverse the same modules. Otherwise the packets will be routed over a external packet-based network such as the Internet. The delay through a packet-based network such as the Internet the delay will vary, depending on distance to the receiver, traffic, capacity and can therefore be difficult to specify.

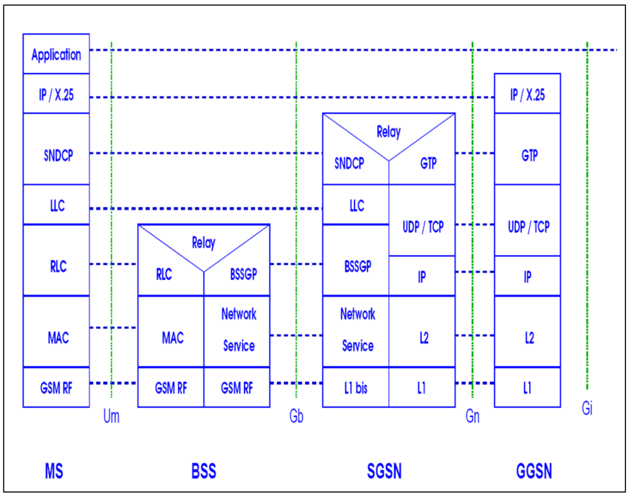


Figure 3.4: GPRS protocol architecture [12]

Processing due to encoding and decoding of voice, music and video and packetization, segmentation, fragmentation and routing of packets also adds additional delay.

The size of the packets is also an important factor which figure 3.5 illustrate. The graph shows that UMTS handles larger payloads much better than the GPRS network.

Delay has been devoted ever more research effort and it is realized that new mobile multimedia applications require low latency to be user friendly. As figure 3.6 shows, for every new packet-based version of cellular network, the end-to-end delay has been reduced, and the aim is that the delay will be reduced even more with new mobile versions. This is essential to achieve if new mobile multimedia applications and services should be successful.

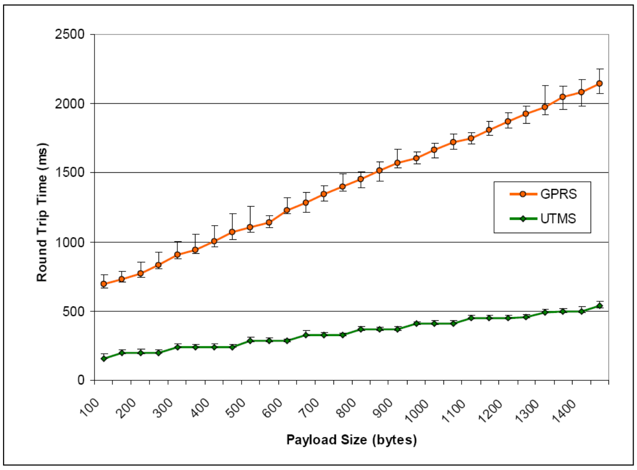


Figure 3.5: RTT measurements for UMTS and GPRS with different payloads[28]

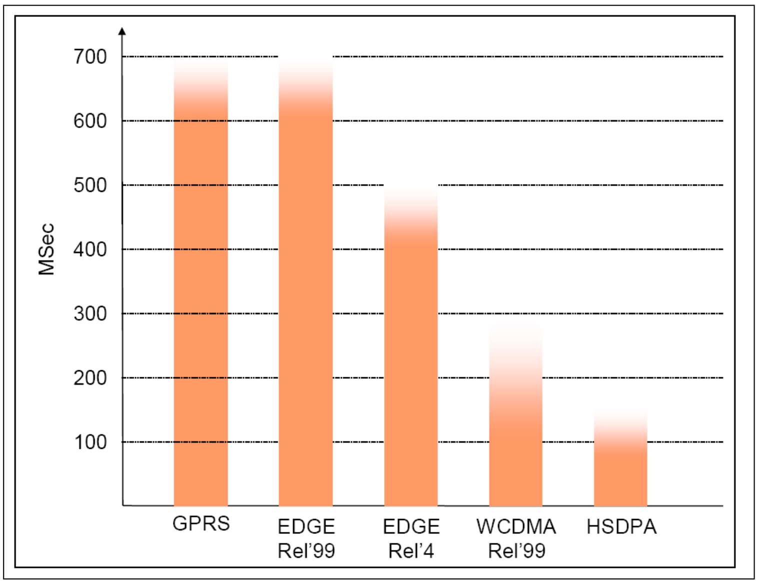


Figure 3.6: Latency for different mobile wireless networks [30]

3.3 Mobile Terminals

There are several mobile terminals with different operating systems and connection options in the market today. The most basic mobile phones has only a keypad, microphone and speaker, but this is changing today. New so-called smartphones and the convergence between PDA's and mobile phones result in highly advanced terminals, capable of doing complex tasks. This section will look more into what options and possibilities customers may choose among today.

3.3.1 Operating System

In the beginning of the 'mobile age' every handset-manufacturer delivered their advanced mobile phones called smartphones with a proprietor embedded operating systems (OS). Due to the ever increasing complexity of new smartphones with 'unlimited' functionality and the joint effort to have open standards, it is a trend today that the manufacturers are merging toward's just a handful different OS. There are mainly three OS' that looks to be dominant in the future of mobile terminals. That is Symbian which is owned by Nokia, Sony Ericsson, Panasonic and Siemens AG. Windows Mobile, which is developed by Microsoft. The third is OS' based on Linux, such a Trolltech's Qtopia [5].

Until just recently the Symbian OS was the dominant operating system for advanced mobile terminals. This is because some of the worlds biggest mobile phone manufacturers such as Nokia and Ericsson are using the OS in their smartphones. Despite the use of the same platform, every manufacturer use their own proprietary graphical user interface.

When Microsoft, which we know as the manufacturer of Windows OS for PC decided to move into the mobile terminal market, it was obvious that they would be a strong competitor. Most PC users are familiar with the Windows OS. It is natural that many users would like to have the same interface and access to the same services and applications on their mobile terminal.

The release Windows Mobile 5 comes in to different versions, one version called Smartphone and the other called PocketPC, illustrated in figure 3.7. The main difference between these versions is that they are fitted to two different types of advanced mobile terminals. The Smartphone edition is more like a regular mobile phone with a keypad and a screen. The PocketPC version is developed for PDA's with a touch-sensitive screen [31, 19].

Linux OS has become very popular for PC, and especially servers which require stable . Now has the trend come to mobile terminals. Different companies are developing and selling handheld devices using Linux as OS. Compared to the two latter OS' Linux-devices are quite rare, but it is reason to believe that Linux will after some time become as popular for handheld devices as it has become for PC's and servers.

Figure 3.7: To the left is the Smartphone version and to the right a PocketPC version of the Windows Mobile operating system.

3.3.2 Connectivity

There is a merging trend that new terminals released on the market comes with different embedded connection options. Earlier the mobile phones had just a voice line connection and PDA's did not have any type of connection possibilities. Now both mobile phones and PDA's has the most elementary connections such as circuit switched voice- and data connection. But upgrades to the cellular networks have provided packet switched data-services which utilize the network capacity better. This capability has made room for developing new services which requires always-on connectivity and relative big volumes of data transfered. The next section will go through different relevant connection possibilities available, with emphasis on data connections.

3.3.2.1 Cellular Technologies

There has in recent years been developed several different cellular technologies. As already mentioned 2G systems like GSM and 3G systems like UMTS have been widely deployed in several regions throughout the world. But there are alternative systems like cdmaOne and CDMA2000 [2] which also have been widely deployed.

GSM and UMTS are originate from Europe, while cdmaOne and cdma2000 originate mainly from North America. But both standards are more widespread in that area. Cellular technology is naturally the most common connection type used in mobile terminals today, but this may change in the years to come.

3.3.2.2 Wireless LAN

There is a increasing trend that new mobile terminals come equipped with a Wireless LAN (WLAN) transceiver, usually based on the IEEE 802.11 standard [3]. The IEEE 802.11 family of technologies has experienced rapid growth, mainly in private deployments. In addition, operators, including cellular operators, are offering hotspot service in public areas such as airports, restaurants and hotels. For the most part, hotspots are complementary with cellular-data networks, as the hotspot can provide broadband services in extremely dense user areas, and cellular networks can provide near-broadband services across much larger areas. Various organizations are looking at integrating wireless LAN service with GSM/UMTS data services, including the GSM Association which is developing recommendations for SIM-based authentication for hotspots and 3GPP which is developing an architecture where WLAN is included as part of UMTS Release 6. WLAN is looked upon as a complementary technology to cellular networks, because it can provide high capacity connection in areas in dense populated areas and cellular technologies can provide connectivity in areas where it is not feasible to deploy WLAN, though with lower data capacity.

3.3.2.3 Satellite Technologies

To provide connectivity in the most waste place on earth, satellite communication is needed. This option in mobile equipment is for people and businesses with special needs such as explorers and ships. The solution is too expensive and not very flexible for ordinary people, who for the most of the time can manage fine with a terrestrial system. The satellite technology has been improved

3.3.2.4 Wi-Max

Wi-Max is a new emerging wireless communication technology standard which may provide high-rate connectivity over relatively long distances. The technology is still quite new, but it is believed that mobile terminals may be equipped with this technology in a near future.

3.3.3 Input and Output Channels

There exist different channels for input and output data to a mobile terminal. This section will describe the most common forms of interaction between human and machine.

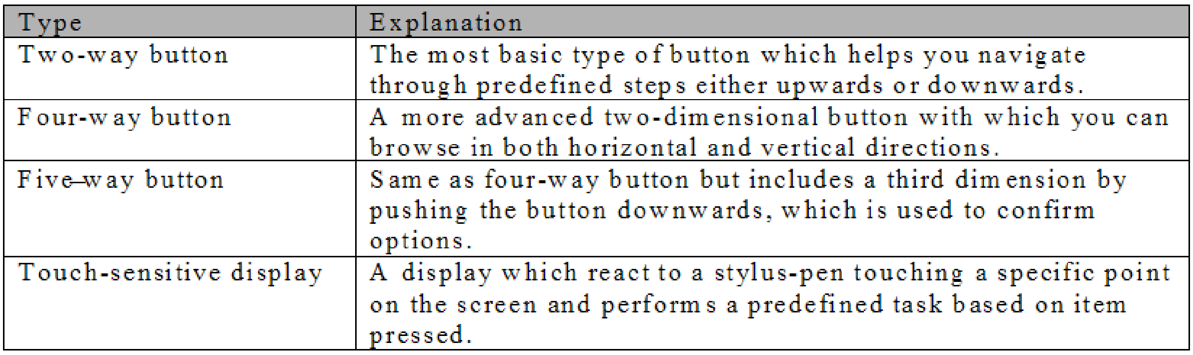
3.3.3.1 Typing

To operate a mobile terminal and be able to use the different services provided, the terminal need to provide some sort of typing methods. The most basic solution is a 12 digit keypad. Additionally the different keys have been assigned at least three letters each. Using these keys to write text can be quite cumbersome and a trial of patience because selecting one letter requires multiple key presses. As a result, different methods for faster and easier typing have been developed. Predictive text technology makes it easier to type text messages on small mobile devices. The technology allows words to be entered by a single key press for each letter and the technology uses a embedded dictionary to predict the word the user is typing.

To embed a full ‘qwerty’ keyboard is another approach to provide typing capabilities on a mobile device. This is sometimes a good solution because many people are accustomed to a regular keyboard. But this involve a compromise, because most people want their mobile device to be as compact and portable as possible, but embedding a full ‘qwerty’ keyboard brings along different size and design challenges.

3.3.3.2 Navigation

To navigate through menus, applications and web-pages on a mobile device different options are available. All devices have a least one dedicated button or function for navigating through different screens. Table 3.6 shows some options for navigating on a mobile terminal.



Type	Explanation
Two-way button	The most basic type of button which helps you navigate through predefined steps either upwards or downwards.
Four-way button	A more advanced two-dimensional button with which you can browse in both horizontal and vertical directions.
Five–way button	Same as four-way button but includes a third dimension by pushing the button downwards, which is used to confirm options.
Touch-sensitive display	A display which react to a stylus-pen touching a specific point on the screen and performs a predefined task based on item pressed.

Table 3.6: Different methods for navigation on different screens on a mobile device

3.3.3.3 Display

The display on a mobile terminal is mostly used for output to the user. But more and more terminals comes with a touch-sensitive display, using a so-called stylus pen to highlight object on the screen. This trend has emerged much due to the convergence of PDA's and mobile phones and to provide better usability operating complex applications on the terminal. The touch-sensitive screen on smart phones also support writing letters which are recognized by the terminal and to create drawings.

Due to the introduction of multimedia applications on mobile devices, the need for larger screens with higher resolution has become evident. But the size and resolution of the display is limited by the consumers 'size-acceptance' of the terminal and what the consumers are willing to pay. The technology inventions are in progress and substitutes to the conventional display is under research such as projected screens, roll-able screens and head-mounted displays.

3.3.3.4 Audio

The auditive mode consist usually of a speaker and microphone. These devices have usually been implemented to use the terminal as a telephone holding the terminal close to the source, i.e. ear and mouth. For a while new 'handsfree' solutions have appeared, in which you can hold your terminal at a arm length distance. The different options are ear piece which is connected to the terminal with a wire or Bluetooth. Another option is to implement a highly sensitive microphone and a loud-speaker into the terminal. One of this solutions or similar is required when interacting with a multimodal interface, because one needs to simultaneous look and operate on the display.

But the microphone and speaker can also be used in other settings to provide both input and output. Communicating with a machine requires speech technologies, which are described in the next section.

3.4 Speech Technologies

Speech technology is an important topic in speech-centric multimodal interaction. In the next section some details about the technologies will be given.

3.4.1 TTS - Text-To-Speech

Text-To-Speech is defined in reference [14] as:

> Text-To-Speech synthesis can be defined as computer-generated speech used to translate electronically stored text information into aural information.

The field has been devoted much research over several decades and the first recognized development in to the area was done as early as 1791 by Wolfgang von Kempelen. Still there is a great potential for improvements and currently the research and development within the area are more extensive than ever.

There are many areas of application for TTS, ranging from simple talking clocks to advanced 3D generated talking heads with synthesis of unrestricted text and simultaneous synchronization of mouth and head movements.

When the first TTS systems became available, specially blind persons were eager to try out the technology, even though the quality of the first systems were quite poor. It is in the area of handicapped people TTS has proved its potential, but also for educational purpose, multimedia and in telecommunications areas are TTS often used. One well known example is related to telephony information services, where the user is met by an virtual assistant telling you where you are and which keys to press to make use of the service.

The TTS process goes through several steps when translating written text into speech. Figure 3.8 shows a high-level overview of the steps performed when translating written text to synthetic speech. A TTS system is divided into two parts, a front-end and a back-end system. In the front-end system the written text is converted into their written-out word equivalents. Then phonetic transcriptions is added to each word and the text is divided into phrases (prosodic units). The combination these processes make up the symbolic linguistic representation which fed as output in the back end which makes up the synthesized speech.

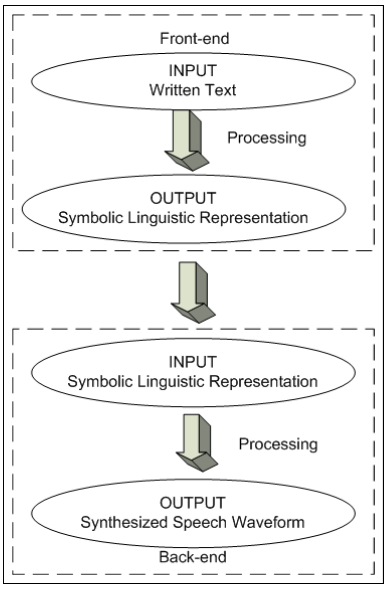

Front-end
INPUT
Written Text
Processing
OUTPUT
Symbolic Linguistic Representation
INPUT
Symbolic Linguistic Representation
Processing
OUTPUT
Synthesized Speech Waveform
Back-end

Figure 3.8: Illustration of the steps in a text-to-speech process.

3.4.2 ASR - Automatic speech recognition

ASR is very broad topic in the area of speech technologies and this section will barely touch the surface of this highly complex process. The reference [33] gives a thorough introduction of the area. The basic idea of ASR is to transform human speech to a sequence of words. These words are understandable for a computer and can, depending on the user context, be used a dictation tool or a commanding tool. The former is used for transcribing text, while the later is used to give a system commands to perform certain tasks with speech. There is many degrees of complexity for a ASR system. The most basic ones implemented on mobile phones, which gives the user the possibility to call up a contact by saying the contact's name. To the more complex ones that understand natural speech and can understand the meaning of longer spoken sentences.

Figure 3.9 illustrate the different steps when analyzing speech. First the speech needs to be converted into a digital representation. Then the computer can run through different procedures to to accomplish the sound-to-word process. It is important that the speech input to the system is as similar as the natural speech as possible. This is due the applied algorithms that is used on the speech. It is a trend that different media are compressed via advanced codecs to generate as little data as possible. This can result in problem for ASR system, because characteristics in the voice that the ASR process uses, can be removed, thus the recognition rate will fall.

It is out of scope for this thesis to go into specific details in each step performed. But the figure illustrate the different procedures and modules that are involved in the ASR process.

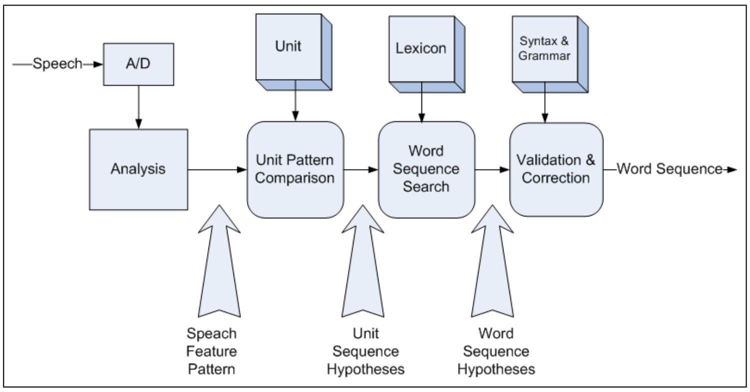


Figure 3.9: Basic components of a speech recognition system [33]

ASR has still not become a everyday application. There have been attempts to introduce the technology in different standard software applications. One example is Microsoft Office, which both provides a command tool, but also as a dictation tool. These tools have not become a ‘killer-app’ among ordinary users. On the other hand certain disabled persons will very much appreciate such tools. Another area of application is in call center systems where the caller is asked to instruct the system using his voice, why he is calling and who he would like to speak to. Based on the spoken information, the system can forward the call to the right human customer service agent.

3.5 Relevant Web Technologies

In the recent years new technologies for designing and developing web-pages has arised. These technologies provides new features and creates possibilities for developing new services and applications on the web. This results in better usability, better mobility, faster and more always up to date features. This section will shortly describe a couple of these web technologies that are relevant to this project.

3.5.1 XML

XML (Extensible Markup Language) has become a very popular standard. As many of the other web standards, XML is worked out within the W3C organization. XML is more or less a generic form of the well known HTML standard. The XML standard has been adopted in several other standards as the reader will discover later.

3.5.2 AJAX

AJAX (Asynchronous Javascript with XML) is a very promising concept in web development. The concept is based on the well known standards; Javascript and XML. Applying both standards in a common interface provide the user with dynamic web pages and give the user a feeling of a more responsive web page by exchanging only small amounts of data with the server. In this way the entire web page does not have to be reloaded each time the user makes a change. This is meant to increase the web page's interactivity, speed, and usability.

3.5.3 SMIL

Synchronized Multimedia Integration Language (SMIL) is an open standard developed by W3C [36]. A SMIL document describes multimedia presentations using the XML and include timing, layout, animations, visual transitions. Media objects that are included in the presentations is referenced by URL's. The standard is very relevant and have been adopted in applications developed for handheld devices. For example, it has been used in the MMS standard, a service implemented in mobile networks to enable subscribers to exchange short multimedia messages comprising text, sounds, pictures and video.

3.5.4 VoiceXML

VoiceXML is a markup language for creating voice user interfaces that use automatic speech recognition (ASR) and text-to-speech synthesis (TTS). It is the W3C's standard XML format for specifying interactive voice dialogs between a human and a computer. The resemblance to HTML is striking. In the same manner as HTML provides a visual interface to the user, VoiceXML provides a voice interface. VoiceXML documents are interpreted by a voice browser in the same way as HTML are interpreted by a visual web browser. VoiceXML has tags that instruct the voice browser to provide speech synthesis, automatic speech recognition, dialog management, and sound file playback.

Chapter 4

System Overview

This chapter gives a general overview of a multimodal system and the MUST multimodal system mentioned in section 2.2. Then some scenarios will be given resulting in some requirements. These requirements will constitute the basis for implementing a mobile multimodal service.

4.1 General Multimodal System

Figure 4.1 shows an simplified overview of a multimodal system. A multimodal system has a server-client architecture. The client part is quite simple comprising basically two parts. The input part which sends user interactions to the server such as voice and pen input and a output part which presents the results received from the server. The server side is very complex, consisting of different modules performing system critical tasks. As figure 4.1 illustrates, the server has a input part which collects different input from the user and forwards it to the Dialog and Interaction Manager (DAIM). The DAIM module process the input and interact with the application specific module to generate a result to the user. The result is forwarded to the output module which make the result of the user query presentable in the form most suitable.

Type of application, illustrated in figure 4.1 that is implemented over the DAIM module is of less importance in this project. The services provided to the users can be applications similar to the ones mentioned in section 2.2.1. It is just imagination that sets the limit of what kind of service one would like to provide using a multimodal interaction system.

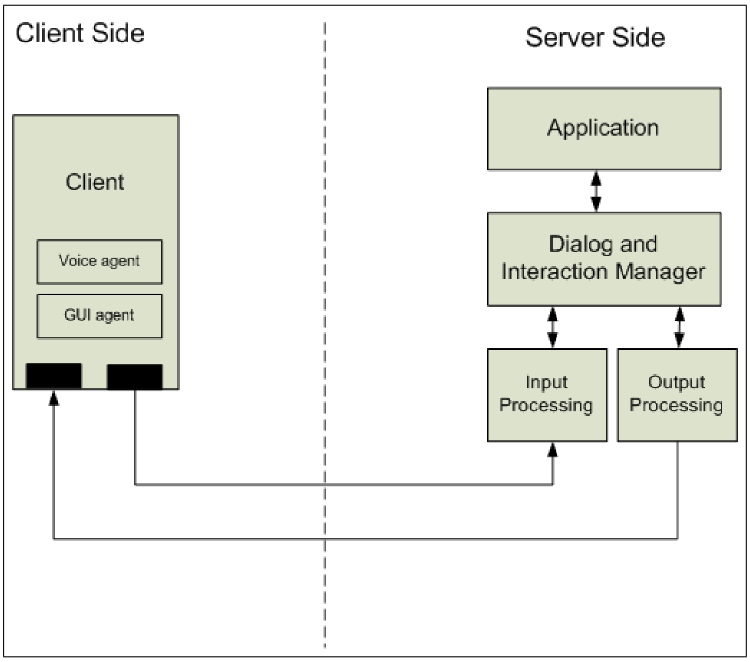


Figure 4.1: Simplified overview of a multimodal system

4.2 The MUST System

This section will describe the MUST multimodal platform and is based on the system description given in reference [37]. The server and client part and the interaction between them will be described. The MUST system was developed for demonstrator purpose, thus it is limited in functionality and usability. Nevertheless, it serves as a good foundation for creating a mobile multimodal service platform.

4.2.1 Server Side

The server part of the MUST system consists of five main autonomous modules (or servers) that communicate via a central facilitator module called a hub, illustrated in figure 4.2. These modules will be described in the subsequent sections.

4.2.1.1 Hub

The different server modules need to communicate with each other to perform certain tasks. To handle these messages a hub is implemented, as figure 4.2 shows. The ingenious with the hub is that it provides modularity to the system. Messages are distributed between server modules according to certain rules based on the service logic. The messages are usually asynchronous

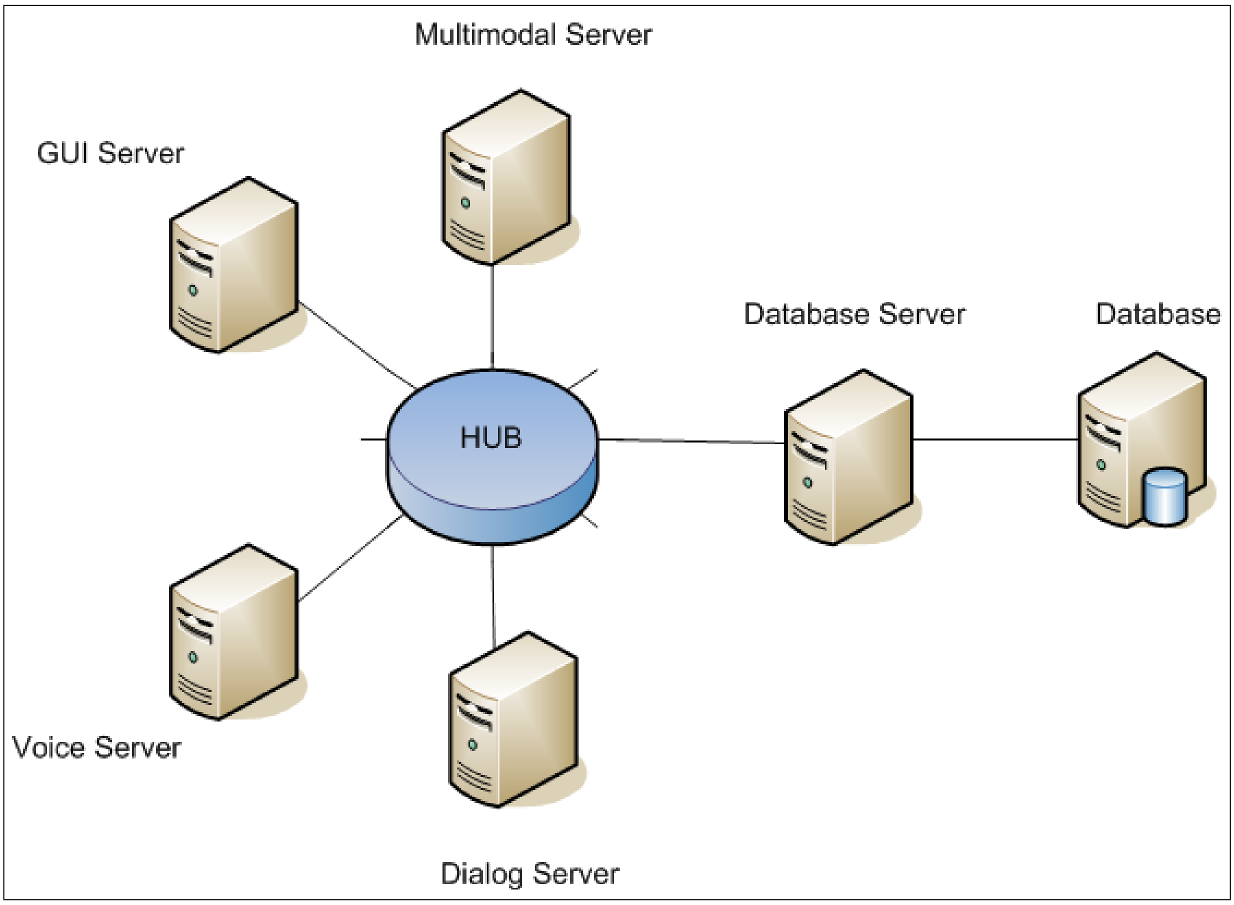


Figure 4.2: Server architecture of MUST

which means that the modules cannot expect to receive a response immediately. A module requiring a certain functionality may pass this job to the hub, and the hub will then know which module to forward the request to. This makes up the properties modularity, distribution and seamless integration of modules which constitute the hub.

4.2.1.2 Voice Server

The voice server is the module that handles the voice modality. Interaction is handled both ways, hence speech input from the user is interpreted by the ASR module and voice output from the system to the user applies the TTS module to construct synthesized voice.

The voice-server supports both packet-switched and circuit-switched voice transmission. The packet-based version is a simple proprietary VoIP solution. It simply copies the audio input and converts it to a standard PCM format and transfer it over a TCP/IP socket connection between client and server.

The circuit-switched version is a standard ISDN solution. It is a straightforward solution and the server requires a ISDN interface connected to a ISDN telephony system. The solution makes it easy to set up a voice connection between the server and the client with a standard mobile phone such as GSM or UMTS.

The voice server also includes a voice activity detection (VAD) module which tries to sort out specific messages given from the user and general noise generated from the surroundings of the user. Because both voice transmission solutions transfer everything that the client device detects, VAD is necessary so that the ASR module does not get overloaded with audio input to interpret.

4.2.1.3 GUI Server

The GUI server handles the visual and tactile modality meaning graphics and text output to the user and the pen-input received from the user, respectively. It act as the gateway between the client and the other server modules.

Pointing input from the user is received and forwarded to the Dialog Server via the Multimodal Server. Based on feedback from a user query, the graphics and text presenting the result is handled by the GUI server. The GUI server uses a web server to display the graphics and text.

4.2.1.4 Multimodal Server

The Multimodal server receives input from both the voice and GUI server, respectively speech and pen click. The inputs collected here are the first step of integrating the two modalities. Within a predefined time window the inputs are collected and forwarded. This means that for simultaneous coordinated multimodal interaction where both pen and speech are applied, these inputs are gathered by the multimodal server and forwarded directly. If only one of the modalities are used, the multimodal server will wait until the time window expire before it forwards the input to the Dialog server.

4.2.1.5 Dialog Server

The Dialog Server also called Dialog and context manager module is the most important part of the multimodal system. The Dialog server receives the user query which may contain both voice and pointing inputs from the multimodal server. Based on the these inputs the Dialog server extract the meaning of the user interaction. Further the dialog manager interact with the database server to generate an answer to the user query. The last step is to present the query response and transfer the information to the user comprising both speech and graphic.

4.2.1.6 Database Server and Database

The Database Server is application specific and should be a general as possible to support all kinds of applications. The Database server act as a connecting link between the Dialog server and the database. It is in the database the application specific information is stored and retrieved.

4.2.2 Client Side

This section will only describe general properties of the client, more details is provided in section 4.4 which explains the client part and development details. Basically the multimodal client comprise a Voice Client and a GUI Client, both incorporated in a standalone software product. The software is developed for the OS Microsoft Windows Mobile 5.0 Pocket PC edition. Thus it can only run on terminals running that OS.

Figure 4.3 shows a logical overview of the different components that the client consist of, namely the Connection Manager, GUI client and Voice Client. The Connection Manager provides a interface between the Voice and GUI module and the network and consequently the multimodal server. The Voice client handles the voice modality, i.e. it receives and forwards voice commands from the user and output synthesized voice from the multimodal server.

The GUI client is somewhat more complex. It consist of a web browser which retrieves web pages containing the graphical user interface and consequently the application provided. It also handles pen-input from the user, i.e. when a user points on a icon on the web page, the coordinates of the pushed icon is collected and transferred to the multimodal server as user input and handled thereafter.

The client communicate with th server using a pre-established ad-hoc WLAN connection[1] using the IEEE 802.11x[2] standard. This is not a mobile nor scalable solution, but it functions well for demonstrator purpose. The WLAN connection is capable of transferring both the GUI data and voice data.

4.2.3 Information and Communication Flow in the System

This section will explain and illustrate the interaction pattern between user, client and server. Figure 4.4 illustrate the information flow between the modules in the MUST system.

[1] A point-to-point connection.

[2] X here refers to standard A, B or G of the IEEE 802.11 standard [3]

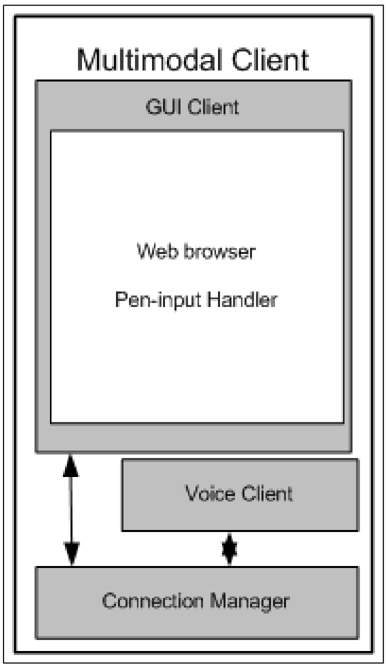


Figure 4.3: Multimodal client architecture[37]

The information flow is based on user queries containing voice and pointing input. These inputs are transferred to their respective server modules. The GUI server register where the user has pointed and the voice server performs a voice recognition and extract the essential meaning of it. Next these inputs are handed over to the multimodal server, which employs a timer mechanism to collect input signal within a specified time window.

Further, when the time window expire or the multimodal server has received a specified maximum of simultaneous inputs it passes these inputs to the Dialog Server. The Dialog Server completes the multimodal integration and based on this process it will try to create a response to the query. To help out the with elaborating a response, the Dialog Server can query the Database Server. The Database Server performs a lookup in the database and returns the result back to the dialog module.

The result is then processed by the dialog manager to create a presentable response to the user. The response is passed over to the multimodal server which splits the information in to different modalities i.e. graphics and speech which are sent out via the GUI and voice server respectively.

As figure 4.4 shows the client needs to set up two logical channels between client and server. One channel for transmission of the voice modality and another channel for data, comprising pen-input, graphic and text output. The communication between client and server is detailed in figure 4.5. Step 1 to 6 is already described above. At step 7 the multimodal server creates a web page that is uploaded to a web server. At the same time a message is sent to the client telling it

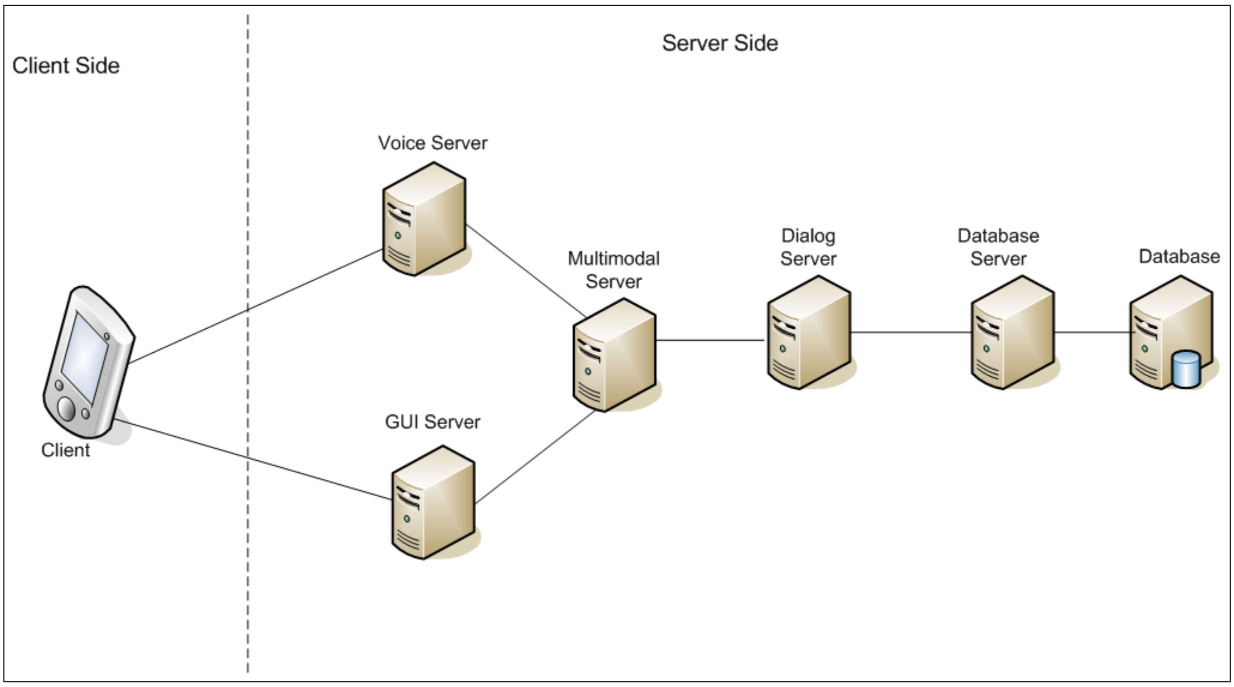


Figure 4.4: Information flow through the system [37]

to download the web page at the given URL. The client which has an embedded web browser send a standard HTTP request to retrieve the web page created. At step 10, the multimodal system transfer the voice response elaborated from the result of the user query. Preferably in the same moment, the web page presenting the visual modality of the result is displayed at the client. The synchronization of these two outputs are as crucial for the user experience as the synchronization of the user input. Because the multimodal interaction is performed in real-time, the different modalities are dependent on each other and must be handled thereafter. This is a important aspect that will be discussed later.

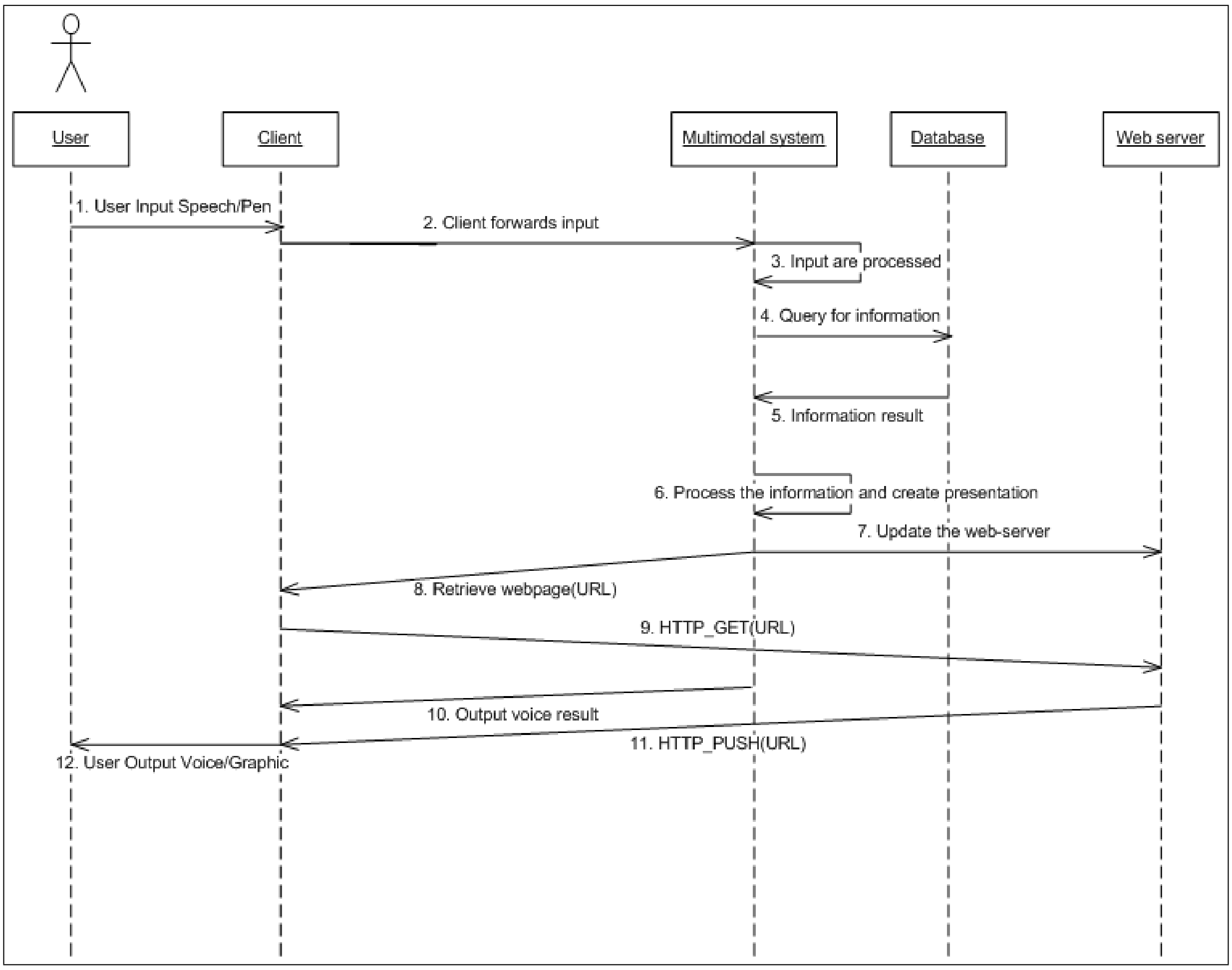


Figure 4.5: Simplified sequence chart describing the interaction in the multimodal system

4.3 Scenarios and Requirements

This section will specify requirements for a mobile multimodal client-server system. The focus will be on the underlying platform providing a multimodal interface to users. But for illustrative reasons a service will be used to better specify requirements to the system. The multimodal platform used is speech-centric and the user can interact using a stylus pen to indicate icons on the screen. This chapter will therefore take these features into consideration when specifying the requirements to the multimodal system.

The scenarios given here will present different situations and different needs that people may have that a multimodal system should provide service for. Based on these scenarios, requirements will be deduced. The requirements are focused on both functional and performance requirements. The requirements will focus on the underlying multimodal platform. I.e. the functionality the multimodal platform makes offer to multimodal applications, thus type of application or service is not considered.

4.3.1 Retrieving bus information

Alice is on her way from job to meet her friend Bob at his place. She is traveling by bus. She has visited Bob earlier, but never directly from work. She is taken the usual bus the central station, but from there she do not know the way further. She wants to use her mobile terminal supporting multimodal interaction. She logs into the bus information service using the UMTS network. Alice feels uncomfortable saying out loud speech instructions when sitting on the bus. Thus she is only using the pen to point the departure and destination addresses on the map and type of information she would like to obtain. Alice thinks it is difficult to remember the travel information given to her, so the bus information is in addition to be presented orally and it is also displayed textually on the terminal.

Requirements:

- The system shall be available over a mobile wireless network.
- It shall be possible to give input to the multimodal service based on user preferences, i.e. voice, stylus pen or both.
- The system shall present the results of a user query based on user preferences, i.e. graphic, text, voice or a combination.

4.3.2 Retrieving bus information 2

Bob suffers from aphasia and can not express whole sentences. He has also a problem with navigating through complex graphical user interfaces on a mobile terminal. Bob wants to know when the next bus leaves from his house to the doctor. He knows where the bus stations is situated on the map of his multimodal terminal. He logs into the bus information service on his multimodal supported mobile terminal. To retrieve the bus information he tells the application “Bus here to here?” while simultaneously he pointing at the two bus stations on the map. The result of the query is then read aloud to him, but also displayed as text on the screen of his terminal. In this way he easier can comprehend the information given to him.

Requirements:

- The system shall be able to receive simultaneous inputs using both speech and stylus pen on a touch-sensitive display.

- The mobile network must support simultaneous transmission of real-time voice and data.
- The transmission of voice must maintain the audio quality such that the ASR module can achieve a high recognition rate.
- The services provided over the multimodal platform shall be easier to use than conventional telecom services.
- The multimodal system shall attract new customers that usually are uncomfortable with operating services on their mobile terminal.
- The multimodal system shall provide responsive services.
- R5 - The multimodal system shall handle the characteristics that the infrastructure provide, i.e delay, data rate etc.

4.3.3 Mobile Operator

MobiTel is a mobile network operator with a license for UMTS. They have just recently opened their network, but they have problems attracting customers. They see that the cost of deploying and operating the UMTS network is high and they can not compete with the other mobile operators on price, especially with operators running a well functioning GSM network. They have realized that their need to focus on providing new and good services. Unfortunately there are very few services in their current UMTS installation that can not be implemented in the GSM network. They discover the multimodal technology, which sets certain requirements to the network, because it needs to transport both real-time voice and data simultaneously, which GSM can not do. They also see that services using multimodal interface can be easier to use and generate traffic in their network.

MobiTel wants to be independent of any mobile manufacturer, so they want their multimodal service to be based on open standards and able run over mobile terminals independent of OS, manufacturer or any other factor that can limit the potential customer market.

To invest in a multimodal system many different services should be available for their customers. MobiTel sees that they can not develop and implement so many services as they wish. Instead they want to just provide the multimodal platform and then can third party service providers implement their specific services on top of MobiTel's multimodal platform.

MobiTel has seen that customers are reluctant to try out new services if the services provided demands much installation procedures from the users. In these cases they will only attract users with great interest in new inventions and probably not attract the average user, suffering from low market volumes.

Requirements:

- The multimodal system shall be based on open standards not restricting it to specified mobile terminals.
- The multimodal platform shall be a generic platform with a possibility to implement all kinds of services on top of it.
- The multimodal platform shall allow third party service providers to offer their services over the mobile operators multimodal platform.
- The client part of the multimodal system shall be simple, requiring minimal installation for the user on the terminal.

4.3.4 List of Requirements

The requirements identified in the last section are put in concrete terms in table 4.1.

Nr	Requirement description
R1	The system shall be available over a ubiquitous mobile wireless network.
R2	The mobile network must support simultaneous transmission of real-time voice and data.
R3	The transmission of voice must maintain the audio quality.
R4	The system shall be able to receive simultaneous inputs.
R5	R5 - The multimodal system shall handle the characteristics that the infrastructure provide, i.e delay, data rate etc.
R6	The user should be able to use the different input channels to the multimodal interface according to user preferences.
R7	The system shall present the results of a user query based on user preferences.
R8	The multimodal system shall be able to attract all kinds of people and users.
R9	The multimodal system shall provide services that are responsive and intuitive.
R10	The services provided over the multimodal platform shall be easier to use than services applying conventional interaction methods .
R11	The multimodal system shall be based on open standards.
R12	The multimodal platform shall be generic making it possible to implement different kinds of applications on top of it.
R13	The multimodal platform shall support a system interface and API that allow third party service providers to offer their services.
R14	The client part of the multimodal system shall be simple to implement, requiring minimal intervention from the user.

Table 4.1: Requirements for the multimodal system

4.4 Implementation Details

4.4.1 Implementation Issues

As indicated earlier the implementation done in this project have been based on the MUST platform developed in the EURESCOM project. The system in it self is very complex, comprising many advanced interacting modules both on the server and client side. The methods used to implement functionality based on the requirements specified in section 4.3 have been influenced by the complexity of the MUST system, hence the system's existing architecture needs to be followed. It is not feasible as regards to time resources, nor does it serve any purpose to develop a whole new multimodal platform from scratch.

The server modules are adaptable and can be changed to conform to the requirements specified, but as said the server part is very complex. There are many interacting processes and internal operations which makes it difficult to intervene with the server side modules. However the client side gives room for implementing functionality that satisfy one or more of the specified requirements.

4.4.2 Conforming the System to the Requirements

To make the multimodal system mobile a different type of connection between client and server needs to be implemented. The connectivity should provide an ubiquitous, cost-effective and viable service. As described in 3.3.2 there exist many different communication options on mobile terminals. But not all communication technologies meet the requirements for a mobile multimodal service. As the reader probably already has assumed, it has been chosen to elaborate the possibilities to use the the European version of 3G, UMTS for handling traffic between the client and server. The subsequent sections will describe which means were required to make the client mobile and adapted for communicating over a UMTS connection.

Ideally, the client software should be as light as possible generating a small footprint. The basic functionalities of the client is to receive input from the user and forward it to the server. These inputs is user highlighting a certain point of the the touch-sensitive screen using a pen and send this coordinate to the server. Simultaneous it should receive voice input and continuously transmit it to the server without any further processing.

Figure 4.6 shows a high overview of the architecture of the client part. To connect the client with the server via an UMTS connection, new functionality was required. It is indicated in each module, methods added or updated compared to the original version using a WLAN connection.

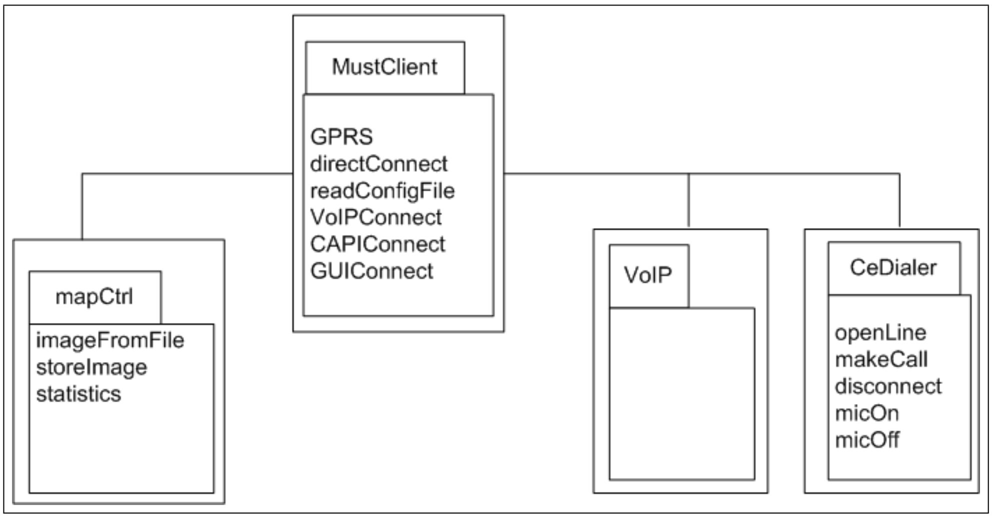


Figure 4.6: Coarse overview of the classes in the client

4.4.2.1 MustClient

MustClient is the main class which controls and utilize the capabilities in the other classes. The following list of items describes the methods added or changed in the class.

- Gprs - Checks whether a GPRS or UMTS packet-switched connection is present and/or establish one according to configurations given by the network operator if not present.
- directConnect - Receives a request from the user to establish a connection with the multimodal server.
- readConfigFile - Reads the enclosed configuration file to determine network communication details (see below for more details.).
- VoIPConnect - Receives a request from directConnect to establish a VoIP connection using the functionality embedded in the VoIP class.
- CAPIConnect - Receives a request from directConnect to establish a circuit-switched connection using the functionality embedded in the CeDialer class.
- GUIConnect - Receives a request from directConnect to establish a GUI data connection with the server.

The emphasize has been on the connection details when altering the client software. To specify connection details, MustClient makes use of a enclosed configuration file. The configuration file specifies:

- Server IP - specifies the IP address of the server to connect to.
- GUI Port - specifies the TCP port to connect to for handling the GUI data traffic.
- VoIP Port - specifies the TCP port to connect to if the voice modality should be performed using the VoIP solution.
- Phonenumber - specifies the telephone number to call if the circuit-switched voice connection should be used.
- CAPI - specifies whether the VoIP solution or the circuit-switched solution for the voice modality.
- Verbosed - specifies whether log data should be gathered during a multimodal session.

4.4.2.2 CeDialer

The CeDialer module is entirely new to the MUST Client software. This module provides telephony capabilities to the client based on a circuit-switched connection. The software is based on a sample distributed in the Windows CE Tools included in Windows Mobile Pocket PC SDK. It has been modified to accommodate the requirements in the multimodal solution. The most important methods will be explained here:

- openLine() - To set up a telephone call the software needs to prepare the telephone capabilities of the terminal.
- makeCall() - The terminal initiate a phone call with a specified telephone number.
- disconnect() - The terminal shut down the telephone call.
- micOn() - Sets the terminal's microphone to active.
- micOff() -Sets the terminal's microphone to passive/mute.

4.4.2.3 VoIP

The VoIP class is responsible for setting up a VoIP connection between client and server. It continuously record the voice and transfer it to the voice-server over a TCP socket connection. The module did not require any changes for it to work over a UMTS connection, hence there is no methods added or changed.

4.4.2.4 mapCtrl

This part is responsible for the GUI modality. It takes care of the pen-input which is transferred through a TCP socket connection to the GUI-server. It also handles the graphical output received from the server. This module did not require much change, but some functionality was added.

- storeImage - saves a copy of a picture downloaded from the server.
- imageFromFile - checks whether a specified picture is stored locally and retrieves it. If not, the picture is downloaded from the server.
- statistics - statistics regarding download rates and delay gathered during a regular multimodal session.

4.4.3 Connection Options

As indicated earlier, the voice-server provides support for both a VoIP-solution and a standard telephony connection using ISDN. Using a 3G terminal that support both standard circuit-switched telephony and packet-switched data, it is implemented support for both standards in the client. To connect the client to the server via a UMTS network, two different schemes are possible and explained in the subsequent sections.

4.4.3.1 An All-IP Solution

As the heading indicates, this solution relies solely on a packet-switched UMTS connection to transfer both voice and data. Both voice and data are converted to data-packets and transferred over a IP-socket established on top of the UMTS data connection. Figure 4.7 shows a sequence chart that describes how the client connect to the voice and GUI server. This is a high-level figure, focusing on the logical communication. Set up of lower layer connections such as a UMTS data connection and IP and TCP connections is not included.

First the user ask the client to connect to the server. The client initiates the directConnect method which checks whether the voice connection should use the VoIP solution or the circuit-switched solution. The VoIP solution is specified and the client initiate the VoIP object to prepare to set up a connection with the given serverIP and voip_Port. Simultaneously the MustClient must initiate a UMTS data connection using the GPRS method. When the terminal has set up a packet-switched connection, it is possible to try to set up a GUI and voice connection with the server.

If either of the connections or both fails, the user is prompted with a connection failed message. If both connections is set up successfully, the client confirms to the user that the client is connected and ready to use.

4.4.3.2 A Combined Solution

The combined solution comprise a circuit-switched connection to transfer the voice and a packet-switched connection to transfer the data such as pen input and graphic and text output. According to appendix A.2, the throughput rate for utilizing both connections is restricted up to 64 kbit/s both uplink and downlink, compared to the the ‘all-IP’ solution where the the throughput rate is according to UMTS standard and practical tests up to 384 kbit/s. But another interesting result found in appendix A.2 is the measured delay for the two options. When applying both connections, the delay for small payloads are measured to an RTT average of 208,5 ms. When only connected with a packet-switched connection the delay is measured to an

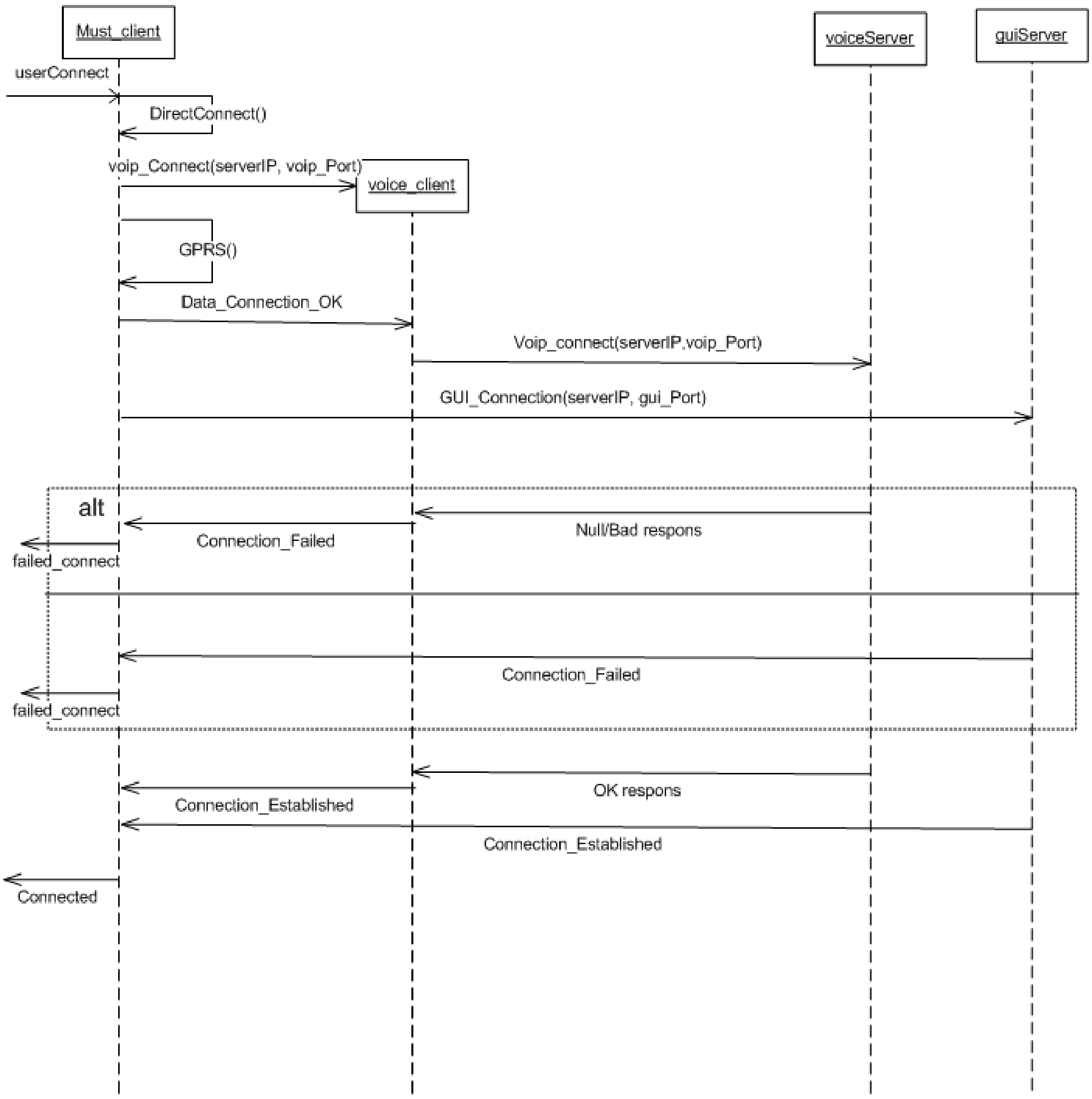


Figure 4.7: Client connect to server using packet-switched voice connection

average of 374 ms, which is almost the double. This is somewhat surprising and results in that the responsiveness of the this solution should be better than the all-IP solution.

The figure 4.8 illustrate the sequence of steps that is performed to set up the voice and GUI connection. The user starts up the multimodal service. The directConnect method detect that a circuit-switched voice connection should be set up. The MustClient ask the CeDialer to set up a voice connection with the given telephone number. The MustClient simultaneously sets up a UMTS data connection using the GPRS method.

The CeDialer sets up the terminals telephony capability and tries to set up a circuit-switched connection with the voice server. The MustClient also tries to set up a GUI connection similar to all-IP solution. If either or both connection set up fails, the user is notified with the message connection failed. If the both connection is set up successfully, the user can start using the service.

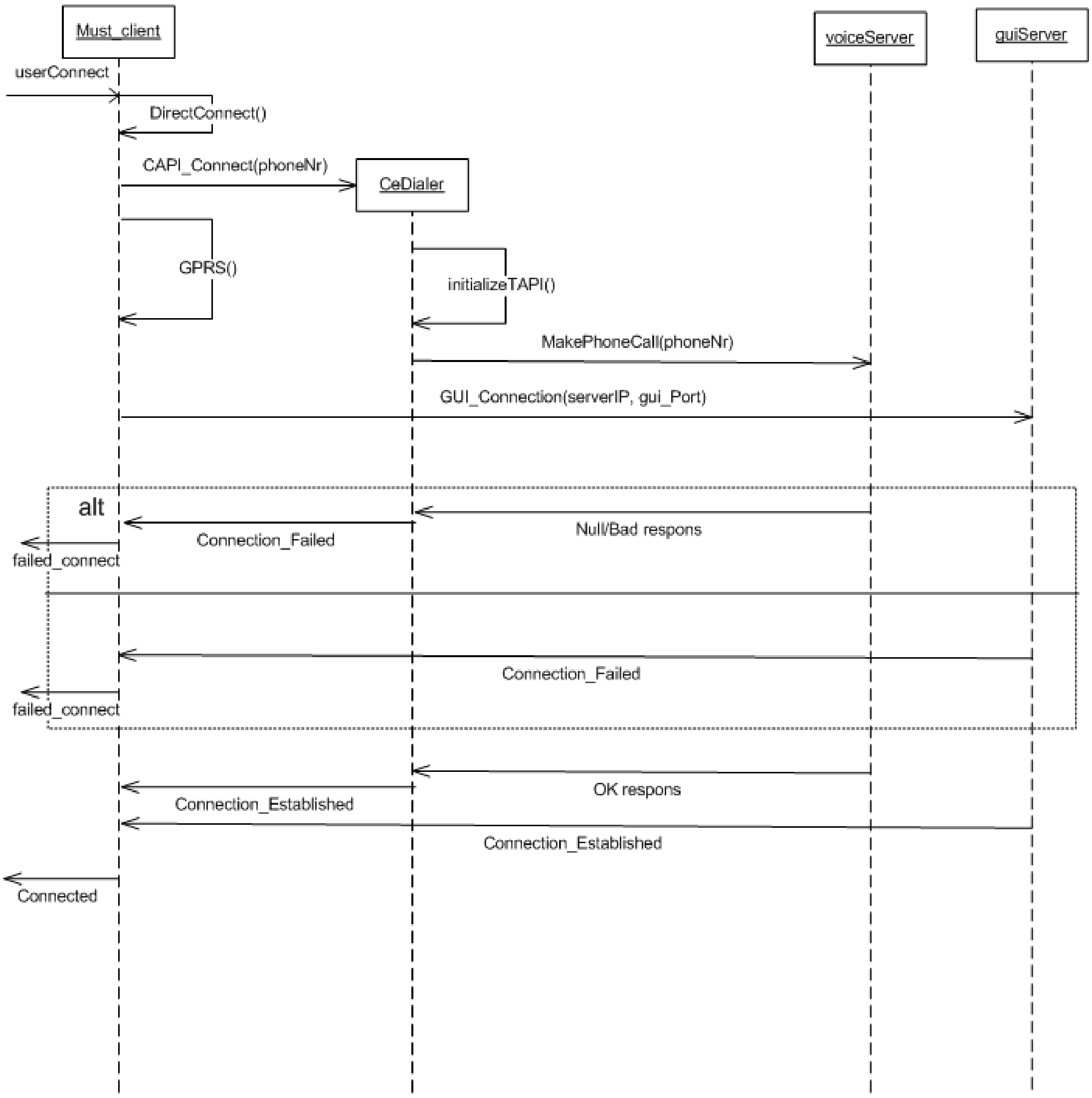


Figure 4.8: Client connect to server using circuit-switched voice connection

Chapter 5

Results

This chapter will present an analyze based on the requirements given in section 4.3. As section 4.4.3 describes there is two different options for connecting the client and server. Thus, it have been performed testing and a general analysis on both solutions according to the requirements. All the requirements specified has not been possible to test directly by using the multimodal service, thus general performance testing on the network has been performed. For requirements that are not possible to do functional nor performance tests, will be answered based on what has been achieved during the implementation phase.

5.1 Analysis of the Requirements

The requirements are segmented into groups, because some of the requirements are congruent. The answers to each requirement group are described thereafter.

5.1.1 Connectivity requirements

- R1 - The system shall be available over a ubiquitous mobile wireless network.
- R2 - The mobile network must support simultaneous transmission of real-time voice and data.
- R3 - The transmission of voice must maintain the audio quality.

For requirement R1, both solutions proposed in section 4.4.3 are based on the UMTS network. UMTS is intended to be a network providing ubiquitous connectivity. For Norway, at this stage of deploying the network, the UMTS network does not provide ubiquitous connectivity. It has not been performed any statistical coverage surveys of the network, but based subjective experience the UMTS coverage was in certain areas and especially in-door viewed as poor to non-existing, even in dense populated areas. The coverage map of Norway in figure 5.1 confirms that the UMTS network at this stage does not provide full coverage. The green areas around e.g Oslo indicates UMTS coverage and it is apparent that it is the dense populated areas that have been the focus of the mobile operators development plans in the first phase. The deployment of the UMTS network is still at an early stage in Norway, thus the coverage will be better as the expansion of the network continues.

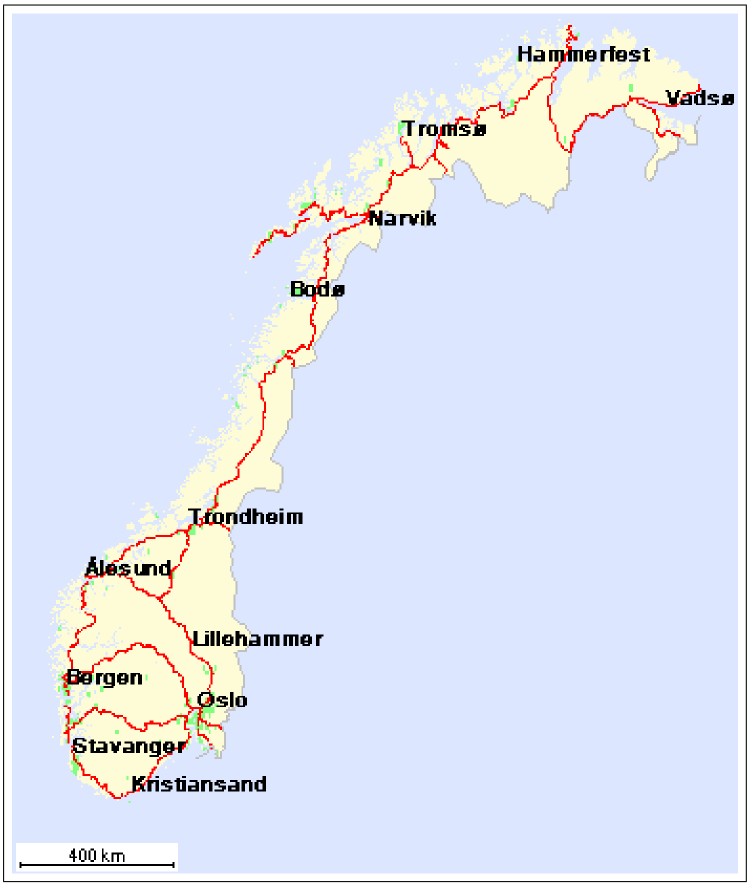


Figure 5.1: Telenor's UMTS coverage in Norway [21]

In areas with UMTS coverage, it was possible to transfer simultaneous real-time voice and data, with both solutions according to R2. But when the UMTS signal strength became weak, the terminal automatically switched to the GSM/GPRS network. This made the all-IP solution unusable because of the poor data capabilities provided by GPRS. For the combined solution the voice channel was still active, despite the handover to the GSM network. But because the terminal is a GPRS class B, the data channel became inactive, thus the GUI modality was no longer working. Another issue is the handover performed within the UMTS network. For the combined solution delay due to handover represents no problem because UMTS CS connection provides good support for real-time services, e.g. standard telephony.

Although UMTS specifies strict QoS parameters for the UMTS PS domain, i.e. it shall support real-time applications specified under the QoS class Conversational, it is difficult to measure if this requirement has been implemented in the current UMTS network. I.e., the handover mechanism described in section 3.1.6.1 is very complex. For the CS domain, experience with handling real-time traffic during handover is good. However, considering handover in the PS domain where the mobile operators currently does not offer any real-time services, the handover delay have not been focused, thus handover can cause problems with real-time traffic in the PS domain. This issue will probably not be dealt with until mobile operators implement the specification for UMTS release 5 which will provide packet-based telephony in their networks.

For what the audio quality concerns for the two solutions as R3 require, it was experienced that the all-IP solution provided poor audio quality. The voice was sometimes garbled and unrecognizable. This is mainly caused by the inefficient VoIP protocol used, but another problem is the relative high latency through the UMTS network. The combined solutions which uses standard circuit-switched telephony, the audio quality was as expected, very good and the ASR module achieved a high recognition rate. Likewise, the speech generated by the TTS module was received with good quality.

5.1.2 Multimodal requirements

- R4 - The system shall be able to receive simultaneous inputs.
- R5 - The multimodal system shall handle the characteristics that the infrastructure provide, i.e delay, data rate etc.

R4 and R5 was tested by using the bus demo application implemented on the multimodal platform. Both solutions was capable of handling simultaneous input. This was tested by employing the feature of coordinated simultaneous interaction described in section 2.1.3.3. But due the poor voice quality of the all-IP solution, the ASR module had problem recognizing the speech command input, thus the system responded that it could not comprehend the user query. For the combined solution the simultaneous multiple input worked fine. But also here it was experienced some difficulties. As appendix A.2.1 shows the delay for the UMTS packet-switched connection can sometimes be long compared to the transmission of voice over the circuit-switched connection, which ensures constant low delay. This resulted in that simultaneous input of voice and touch-input data arrived asynchronous at the multimodal server. Thus the inputs was handled sequentially, and the multimodal dialog manager had problems understanding the user queries because the inputs was complementary and gave no meaning as a single inputs.

This leads us to R5. The multimodal system is solely dependent on the underlying infrastructure and certain mechanisms must be implemented to make the system robust and handle varying network conditions. Though the UMTS network specifies certain QoS requirements described in 3.2.1 which set upper and lower limits on different network performance properties. But it was experienced, as the delay measurement test performed in appendix A.2.1 shows, that the maximum RTT can be up to almost 2.5 seconds. The size of the time window for gathering multiple inputs in the multimodal server described in section 4.2.1.4 can be manually decided. By extending the time window, multiple inputs with different delay through the network can be handled. But it is required a consideration, because longer time-window will result in slower responsiveness of the system and it can occur that sequential inputs that should be handled one by one can be handled as two complementary input resulting in a wrong interpretation of the request. The time window size must be optimized according to the network properties.

5.1.3 User requirements

- R6 - The user should be able to use the different input channels to the multimodal interface according to user preferences.

- R7 -The system shall present the results of a user query based on user preferences.

In both solutions it is implemented the possibility to turn of the microphone. This feature removes the voice modality input, and the user can focus on using the touch-input to interact with the multimodal service. The user also has the possibility to avoid using the touch input, and only give speech commandos to the system. At last as mentioned earlier, the user can also use both input simultaneously, though with some limitations as given in section 5.1.1. By testing the multimodal applications, it was confirmed that all three interaction methods worked fine in both the all-IP and the combined solution.

Functionality for indicating how the user wants the response from the server to be presented was not implemented, though it was no reason to test it. The multimodal server outputs both graphics and voice, regardless of the result presented. This is static set in the multimodal server and additional implementation on both the server and client side needs to be performed.

5.1.4 Usability requirements

- R8 - The multimodal system shall be able to attract all kinds of people and users.
- R9 - The multimodal system shall provide services that are responsive and intuitive.
- R10 - The services provided over the multimodal platform shall be easier to use than services applying conventional interaction methods.

The concept with multimodality in this context is to make it easier for people to use mobile services. The result of this is that mobile services will be more popular and thus attract new users. It is difficult to test if R8 will be fulfilled. To get indication whether this requirement can be fulfilled, a public survey needs to be performed. For R9, this requirement is somewhat related to the results presented in section 5.1.2. Based on the findings in appendix A.1 the responsiveness can in the best case said to be varying. Figure A.2 shows that it can take from about 0.5 seconds to nearly 12 seconds for the combined solution and up to 7 second for the all-IP solution to update the screen of the client. The results is poor and indicate that it is still a lot of work to be done.

It was realized that the multimodal application used the same pictures several times in the GUI modality. To avoid the client requesting the same pictures from the web-server each time the GUI should be updated, a caching mechanism was implemented as described in 4.4.2.4. The aim with the caching function was not to decrease the amount of data transferred, because UMTS provides sufficient data rates for the data traffic generated in this implementation of the multimodal service. The problem is the delay end-to-end in the UMTS network. When requesting a picture from the server based on the downloaded HTML file, the request needs to traverse the UMTS network before the picture is sent back. This adds additional latency for updating the display. By storing the picture locally on the terminal, later retrieval of these pictures will then remove much of the latency generated by the network. The responsiveness was experienced to be better based on trying the service, though it was not done a quantitative analysis showing the improvements. The surprising result discovered and presented in appendix A.2.2 that when using multiple bearers, i.e. both a circuit-switched and a packet-switched connection as for the combined solution, the data rate of the packet-switched connection is reduced to a maximum of 64 kbit/s. This however poses no problems for the responsiveness of the service. The data transferred, constituting the visual modality is not more than it can be handled adequate by a 64 kbit/s connection.

Based on my own experience trying the multimodal service with the bus information service, the multimodal interface was found to be easier to use than trying out a similar bus information service on conventional WAP, according to requirement R10. This is of course subjective considerations and there should be performed a public survey to get reliable results. But there are research projects devoted to this issue. E.g. in reference [29] there was performed some research into this area and and based on experiments they found that a speech-centric multimodal interface was not intuitive for new users and that training would be required for the users to exploit the benefits that multimodal interaction methods provide.

5.1.5 General requirements

- R11 - The multimodal system shall be based on open standards.
- R12 - The multimodal platform shall be generic making it possible to implement different kinds of applications on top of it.
- R13 - The multimodal platform shall support a system interface and API that allow third party service providers to offer their services.
- R14 - The client part of the multimodal system shall be simple to implement, requiring minimal intervention from the user.

The requirements R11 - R14 are very comprehensive, thus have not been devoted much focus in this thesis. All four requirements require extensive development and implementation work on the server part of the multimodal server. As mentioned earlier, this has not been a prioritized target, much due to the fact that the server part structure is very complex and much time and effort would be allocated just to get full knowledge of the internal processes. The multimodal platform developed in earlier project has to some degree used open standards such as XML, but also proprietary server modules such as the TTS and ASR module, thus R11 is not fulfilled. R12 is neither fulfilled, the demo application used is closely built into the server modules. Requirement R12 and R13 are somewhat congruent, and a fulfillment of R13 would to some degree fulfill the requirement of R12. For requirement R14, although it is directed to the client part of the multimodal system, the client and server part are so closely related that to fulfill this requirement would require much implementation work on the server part. The requirements will be discussed more in chapter 6.

Chapter 6

Discussion

The discussion will concentrate on the client part and the communication between client and server of the multimodal system. As mentioned earlier, the server part of the multimodal system is very complex and to make considerable changes to it will require much time. Instead the focus has been to fulfill as many of the requirements specified. And many of the requirement was possible to fulfill without doing intervening with the server part.

6.1 General considerations

Based on the results given in chapter 5 it was experienced that it is possible to provide a speech-centric multimodal interface over a UMTS network. But there are still many issues to consider. At the moment, the combined solution provided the best usability, mainly because the voice quality is handled better in the CS domain than in the PS domain. However, this situation will change when the mobile operators start deploying UMTS release 5 in their network. UMTS release 5 will as described earlier be based on a all-IP infrastructure, hence voice will be fully supported in packet-switched domain.

Although the VoIP protocol used to transmit the voice between client and server is not optimized for a mobile wireless network with high latency and restricted capacity, it is a fact that there are no mobile operators that offers VoIP service over their mobile wireless network today (with reservation, to the extent of the authors knowledge). Although some operators have announced that they will release VoIP service in the near future [34]. It should be said that the characteristics for most deployed mobile networks today are not adequate for such service, especially the delay and the handover in the PS domain is causing problems for supporting VoIP.

6.2 Scalability

The scalability for the two solutions proposed are quite different. For the combined solution which require a CS voice connection can be considered as an inefficient solution. This is because a multimodal service will probably have a similar traffic pattern to normal web browsing, i.e. data are transferred sporadic and not continuously, comparable to a user interacting with the web-browser. The user will probably not be commanding the multimodal service continuously, thus a dedicated connection should not be necessary.

For the all-IP solution, this is a fare more resource friendly alternative. The solution requires transmission resources just when there are multimodal input comprising speech commands and stylus-pen and system output data to transfer, leaving the resources free for others to use when the user are busy reading or viewing results generated from the multimodal server.

6.3 Reliability

The way the multimodal solution is implemented today it is dependent on UMTS coverage. Unfortunately, although the mobile operators are building their UMTS network in huge scale, it will take years before the coverage is as good as GSM is today. It is said that it is unrealistic to achieve 100% UMTS coverage, at least in countries with a topology similar to Norway. As figure 5.1 in chapter 5 shows, it is only the dense populated areas that have UMTS coverage. This practice is cost-effective, covering most of the Norwegian population. But as soon as the user moves outside these populated areas, the mobile user needs to depend on other mobile networks such as GSM, GPRS and EDGE. In the context of the solution proposed, this poses a problem, i.e. it is currently just UMTS that provides the necessary capabilities required for a mobile multimodal service.

For the all-IP solution, using the GSM network, this would require to use either the GPRS or preferably the EDGE packet-switched system. These network has not been tried out, but according to data presented in chapter 3, these packet-switched system does not provide sufficient real-time capabilities. E.g. the end-to-end delay as figure 3.6 shows, the end-to-end delay is at least 500 ms for both GPRS and EDGE, which is not sufficient for real-time services.

For the combined solution, using the circuit-switched connection for voice is currently not possible in the GSM network. As of today there does not exist any GSM terminals that support simultaneous circuit and packet-switched connections, i.e. GPRS class A as described in section 3.1.6.3. Though it has been proposed technology that can enable this feature in efficient and cost-effective way, such as the DTM technology mentioned in the same section. What it comes down to, is that the reliability of a mobile multimodal service is first and foremost dependent on the UMTS coverage. Other reliability issues in context of terminal and software is out of scope for this thesis.

6.4 Availability and openness

This section will consider in general how easy the multimodal solution can be rolled out in the market and which prerequisites that are necessary so mobile operators can offer the multimodal service and that the generic user can apply it. The multimodal interaction elaborated in this thesis set certain requirements to the terminal. First of all the terminal must support 3G. As mentioned the 3G deployment and spreading have been quite relatively slow, but it looks like this trend will turn when the mobile operators put more effort into marketing the 3G network. Another requirement that most current mobile phones in the market does not support is a touch-sensitive display. This feature is mostly reserved for the more expensive business and converged PDA/mobile phone terminals. But a trend today is that the display on the mobile phone is getting bigger, owing to the merging features of being able to play and record films and take pictures with the mobile phone. This can contribute to mobile phone manufactures see the possibility to include a touch-sensitive display. An alternative to touch-sensitive screen can be the touchpad technology often found on laptop computers to steer the mouse pointer.

The multimodal client is today based on standalone software for a Windows Mobile Pocket Edition terminal. As mentioned in section 3.3.1, the Windows Mobile platform (WMP) is gaining increased popularity, especially in the business segment. Nevertheless, currently in the overall mobile device market, the WMP has relative small market share. Thus it restricts the market potential for the application. This does not apply to requirement R11, which require the multimodal system to be based on open standards. To make the multimodal client part interoperable with different kind of terminals a whole new approach should be used, this will be described in section 6.6. Concerning the multimodal server part, all server modules are run on a standard of-the-shelf equipped Windows XP computer. Most of the modules are programmed in Java, which make it easy to deploy the server part on any other OS.

6.5 Usability

The aspect of usability is very important, because the multimodal interface basis of existence is to provide better user friendliness. The usability deals with the intuitiveness and responsiveness of the interface, but also that the results from the service are relevant and accurate.

In context of intuitiveness this is somewhat dependent on the design of a specific multimodal application. But as reference [29] indicated, it will probably be a transitional phase to get used to a multimodal interface and take advantage of the possibilities that multimodal interaction provides. The responsiveness is close related to the underlying platform and infrastructure and how communication between client and server is handled. The way it is handled in the proposed solution additional delay is added. As figure 4.5 in chapter 4 shows, when updating the screen of the client, several request response pairs are required. Steps 8, 9 and 11 could be optimized. Instead of asking the client to retrieve a specified URL, the server could instead instruct the web-server to push the new page to the client. Such optimizations methods could be done to reduce the number of messages transferred over the UMTS network, decreasing the total delay and increase the responsiveness.

To generate exact and relevant results based on user queries, it is important that the user queries are perceived correctly by the server. The ASR module is a critical part in this process, because its task is to comprehend and translate what the user commands the service to do. Thus the ASR needs to give logical results to be further processed in the dialog manager. The ASR is dependent on good speech quality to achieve a good recognition rate, as indicated in section 3.4.2. In this context, the combined solution is preferable because it provides very good transmission of voice. The all-IP solution, as described earlier suffers from from speech quality, thus the recognition rate is not adequate which results in low usability.

Though with a more robust and efficient voice codec, suited for packet-based transmission, the all-IP solution can be sufficient. Research done by IBM can help reducing the problem with low recognition rate using ASR when the voice signal needs to be compressed to be transferred effectively over a low bit-rate channel. IBM has proposed a speech compression algorithm that provides sufficient characteristics in the voice signal to recognize the words even though the bit rate is low and suitable for transferring over packet-based wireless networks [4].

6.6 Possible Improvements

The requirements R11 to R14 was not elaborated in this thesis, but considerations revolving them will be made here. The multimodal platform should rely on open standards. This is not the case for the solution presented in this thesis, but it will be described some possible solutions for enabling this.

The client part of the multimodal system is based on proprietary software. Although it looks like the Windows Mobile OS will become popular, the market for those mobile phones are small. As described in section 3.3.1 there are three OS' that seem to be dominant within the mobile terminal world. This initially restrict us to only develop three software clients, one for each OS. Unfortunately, it is not that simple, because every manufacturer of each OS' implement their own proprietary solution, which more or less makes each mobile terminal platform unique. Thus, another approach is required to easily roll-out clients that support most mobile in the market.

One possible solution is to develop the client application using Java technology or the more specific J2ME version [20] for, which many different mobile devices support today. By developing in J2ME, the client can be installed on several different types of mobile terminals easily. Though it has not been elaborated whether J2ME provide the necessary functionalities to enable a multimodal interaction service on a mobile terminal.

A probably more interesting solution is to develop the multimodal system based on web-technologies. The AJAX technology introduced in section 3.5.2 presents interesting features that could be used. The GUI part of the multimodal solution is already partly based on web-technology, i.e. the client is requested to download a new web page when the graphical interface shall be updated due to a response from the server. By using AJAX, parts of the GUI can be updated locally without the need to download a full web-page each time an interaction or service update is done. The AJAX technology provides the possibility to continuously transfer small segments of data to update the GUI, making it more dynamic. This can be ideal when using the UMTS network, because the AJAX technology will continuously transmit data back and forth between client and server, thus making the delay in the network more or less invisible for the user.

Basing the multimodal system using web-technology, it can be avoided that the user is required to download and install a dedicated software client to the mobile. Instead, the user can just apply the built-in web-browser and sign into a web-page providing the necessary properties for a interacting with a service based on a multimodal interface.

By using web technology based on W3C standards, one can ensure platform independence. Web browser developed for different operating systems will provide a interface between the multimodal application and the underlying platform. It is believed that more and more mobile terminals will have either an embedded web-browser or the possibility to install a web-browser that supports the W3C standards. Due to the increased research activity within the area of multimodality new open standards and requirements to ensure interoperability between manufacturers of multimodal systems have been elaborated. Among them are the VoiceXML standard, requirements worked out by W3C such as SMIL referenced in section and standards elaborated by OMA described in section 2.3. Developing a whole new multimodal system these open standards should be adopted.

The multimodal platform should consider the underlying infrastructure, optimizing methods and interaction between client and server to make the system more responsive and streamlined. Although the combined solution is the most reliable and user friendly solution today, work to make the all-IP solution reliable is necessary for the multimodal interaction method to be a viable solution released on the market.

6.7 Future prospects

The multimodal interactions illustrated in this project has typically been human-machine interactions. But there is no obstacle for using some of the research effort elaborated in this thesis in the context of a multimodal human-to-human interaction. For example when speaking to a human service agent of some kind, it could sometimes be nice to share some information. Typically for a travel agent service when ordering a hotel room over a telephone call, a nice feature would be that the travel agent could share pictures of the hotel and room with the customer. In this way it could be easier for the customer to decide which to hotel to go for. This is a simple example from the private market, and it is probably innumerable other similar services that could employ the multimodal feature.

Equally important is the business-to-business segment where much of the cooperation, agreement and general dialogs between providers and customers are done over telephone. Having the possibility to display and illustrate the products and services and have a form of application sharing during a telephone call would probably be very helpful.

The concept of multimodal interaction does not restrict itself to human input. The input to the multimodal system presented, the interaction with the system is in terms of human input such as pen or speech, but there is no obstacle for applying other types of input to the system as well. E.g., the demonstrator service used in this project, a bus information service, it could be nice if the terminal and application could locate where the user is, thus highlight the relevant area immediately when user start the service. Location info can either be fed into the terminal using location information stored in the network or apply an embedded GPS receiver in the terminal. Other input to a specific service can be environment sensors such as thermometer, barometer or noise-sensors. In this way, the service will know more about the context of the user and can adjust the provided service according to the settings where the user is situated.

Chapter 7

Conclusion

The thesis has elaborated a solution for a mobile speech-centric multimodal interaction interface. The solution consist of a server part handling the logic of the multimodal system and a client implemented on a 3G terminal which provides the interface between the multimodal service and a human.

Based on a multimodal platform, functionality has been added to create a mobile version enabled over a 3G network. To exploit the capabilities in the UMTS network, actually two different solutions has been elaborated. A combined solution which constitute a circuit-switched voice channel and a packet-switched data channel and an all-IP solution which handles all traffic over a packet-switched network. The combined solution can be viewed as a reliable, but less effective solution, while the all-IP solution is more future oriented, though introducing some challenges, such as poor voice quality.

The results gathered during the work with the thesis present several interesting points. First and foremost, the possibility to employ more than one modality when interacting with an automated service via a 3G-terminal is a interesting aspect. To employ both speech and a touch-sensitive display either simultaneously or sequentially can provide better user friendliness. Multimodal interaction can contribute to make things easier to explain for the other party or to faster and easier express to a specific service its intentions, i.e. that a user wants to get information about any area of interest. Such interactions methods can actually be done over a mobile terminal, using a publicly available wireless network, setting minimal restrictions on where the user resides at every given moment. Another interesting point is that this new way of interacting with a mobile terminal also can help to include more of the population. People that currently do not feel comfortable with how the interaction must be performed or people with certain disabilities that restrict them from using conventional mobile services can achieve much making the everyday easier if mobile multimodal interaction is realized as a service to the customers.

Bibliography

[1] Opplysningen 1881. *Mobilsøk 3G (Phonenumber search 3G*. 2006. URL:http://www.1881mobil.no/mobilsok3g.html.

[2] 3GPP2. *3rd Generation Partnership Project 2*. 2006. URL:http://www.3gpp2.org/.

[3] IEEE 802.11. *The Working Group Setting the Standards for Wireless LANs*. 2006. URL:http://grouper.ieee.org/groups/802/11/.

[4] IBM Research Alex Sorin. *Rcognition Cmpatible Voice Compression*. 2006. URL:http://www.haifa.il.ibm.com/projects/multimedia/recovc/.

[5] Trolltech AS. Qtopia. 2006. URL:http://www.trolltech.com/products/qtopia.

[6] S. Baudet, C. Besset-Bathias, P. Frêne, and N. Giroux. Qos implementation in umts networks. 2001.

[7] Lou Boves and Els den Os. Multimodal multilingual information services for small mobile terminals (must). 2002.

[8] H. J. Charwat. Lexikon der mensch-maschine-kommunikation. 1992.

[9] Dinside. *Dinsides surfometer (med Java) (Internet throughput rate test)*. 2006. URL:http://www.dinside.no/.

[10] Luis Almeida et al. The must guide to paris implementation and expert evaluation of a multimodal tourist guide to paris. 2002. URL:http://www.korfint.no/ingunn/publications/IDS2002_MUST.pdf.

[11] EURESCOM. P1104 must - multimodal, multilingual information services for small mobile terminals. 2003. URL:http://www.eurescom.de/public/projects/P1100-series/p1104/.

[12] Xiaoyan Fang and Dipak Ghosal. Analyzing packet delay across a gsm/gprs network. 2003.

[13] University of Wisconsin-Madison Gregg C. Vanderheiden. *Universal Design. What It Is and What It Isn't.* 1996.
URL:http://trace.wisc.edu/docs/whats_ud/whats_ud.htm.

[14] Per Olav Heggtveit. An overview of text-to-speech synthesis. *Telektronikk 2.03*, 2003.

[15] IBM. *Why IBM? - Leadership in multimodal.* 2006.
URL:http://www-306.ibm.com/software/pervasive/multimodal/.

[16] Magne H. Johnsen and Knut Kvale. Improving speech centric dialogue systems - the brage project. 2005.

[17] Kirusa. *Mltimodality in motion.* 2006. URL:http://www.kirusa.com/.

[18] Knut Kvale and Narada Warakagoda. Speech centric mobile multimodal service useful for dyslectics and aphasics. 2005.
URL:http://www.iet.ntnu.no/projects/brage.

[19] Microsoft. Windows mobile 5. 2006. URL:http://www.windowsmobile.no.

[20] Sun Microsystems. *Java 2 Platform, Micro Edition (J2ME).* 2006.
URL:http://java.sun.com/javame/.

[21] Telenor Mobile. *Telenor Coverage Map for UMTS.* 2006.
URL:http://telenormobil.no/dekninginnland/index.do.

[22] Nuance. *Xmode Multimodal System.* 2006.
URL:http://www.nuance.com/xmode/.

[23] Ulf Olsson and Mats Nilsson. Combinational services - the pragmatic first step toward all-ip. 2003.

[24] OMA. Multimodal and multi-device services requirements. *Candidate Version 1.1 (Work in process*, November 2003.

[25] Mark Pecen and Andrew Howell. Simultaneous voice and data operation for gprs/edge:class a dual transfer mode. *IEEE Personal Communications*, 2001.

[26] Third Generation Partnership Project. Quality of service (qos) concept and architecture. 2002.

[27] Third Generation Partnership Project. *3GPP*. 2006. URL:http://www.3gpp.org.

[28] Peter Reichl and et al. Project wisqy: A measurement-based end-to-end application-level performance comparison of 2.5g and 3g networks. 2005.
URL:http://userver.ftw.at/ reichl/publications/WTS05.pdf.

[29] John Rugelbak and Kari Hamnes. Multimodal interaction - will users tap and speak simultaneously. *Telektronikk 2.03*, 2003.

[30] Peter Rysavy. Data capabilities: Gprs to hsdpa. 2004.
URL:http://www.3gamericas.org.

[31] Arne Søiland. Microsoft tror ikke på internett-tv (microsoft does not believe in internet tv). 2006.
URL:http://www.computerworld.no/index.cfm/toppsak/artikkel/id/59514.

[32] Skype. *The whole world can talk for free.* 2006. URL:http://www.skype.com/.

[33] Frank K. Soong and Biing-Hwang Juang. Speech recognition - a tutorial and a commentary. *Telektronikk 2.03*, 2003.

[34] Telecoms.com. *3 VoIP with and without Skype*. 2006.
URL:http://www.telecoms.com/(Requires registration).

[35] Trafikanten. Public transport information service. 2006.
URL:http://www.trafikanten.no.

[36] W3C. *SMIL*. 2006. URL:http://www.w3.org/AudioVideo/.

[37] Narada Dilp Warakagoda, Jan Eikeset Knudsen, and Anders Smeby Lium. Implementation of simultaneous coordinated mulitmodality for mobile terminals. 2005.
URL:http://www.iet.ntnu.no/projects/brage.

[38] GSM World. *3GSM*. 2006. URL:http://www.gsmworld.com.

[39] Oliver Yu and Shashank Khanvilkar. End-to-end dynamic adaptive qos provisioning over gprs wireless mobile network. 2001.

Appendix A

Performance of System and Network

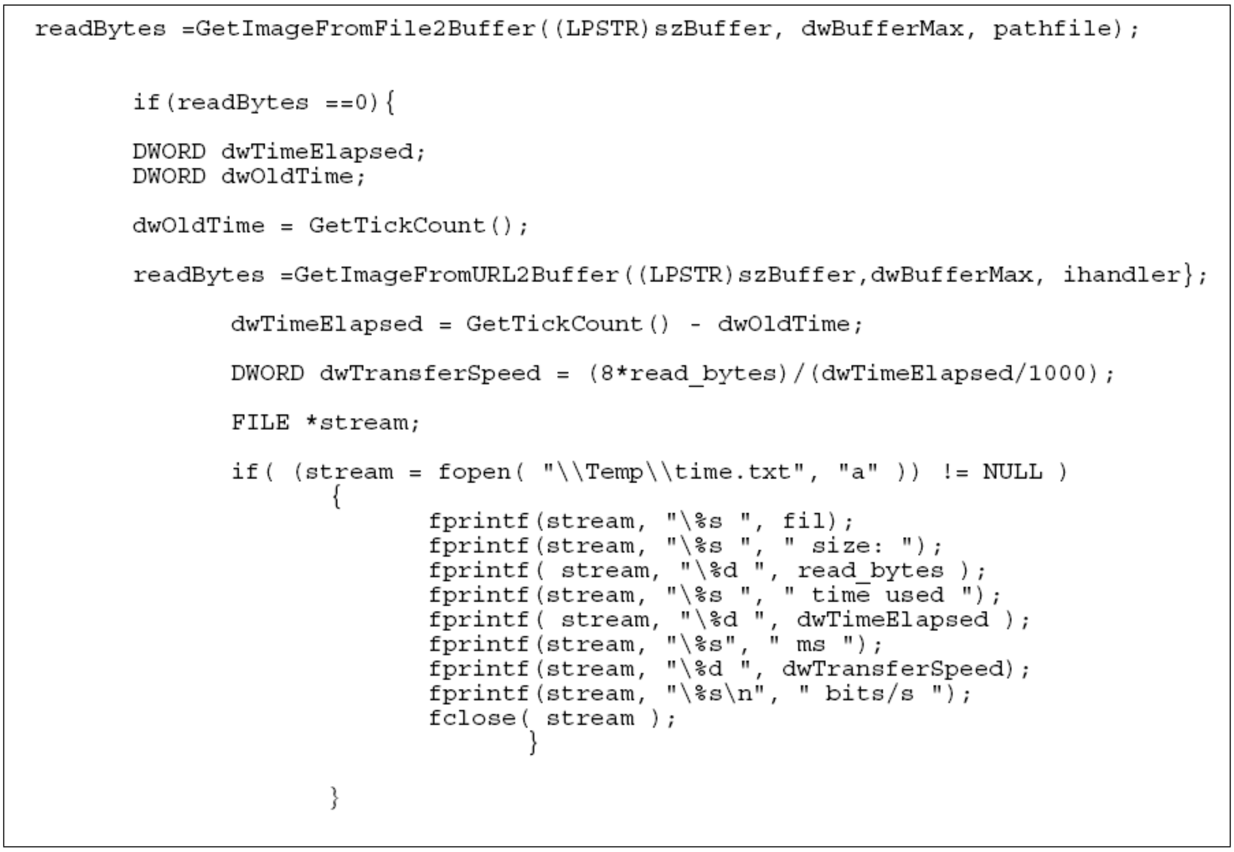
This chapter will present some results from trying out the multimodal system using a 3G terminal. It serves no purpose to do a public survey on the system at this stage. For one, this has been done in other tests such as referenced here [29]. Another reason is that the system based on a mobile connection is not mature enough and will not result in reasonable outcomes. The chapter will also present quantitative result from statistics gathered during testing.

A.1 Practical testing

This section will provide statistics gathered during normal use of the multimodal demonstrator service. Today's people are used to things happens immediately after a request. People are impatient and the threshold for dropping a service is low if the service is slow. Hence, it is important to generate a quick response to a user query. Through the statistic method implemented in the mapCtrl class some quantitative results is collected. The user experience of the system is strongly connected to qualitative results, such as user-friendliness and responsiveness, but with the statistic method a way to get quantitative results are enabled.

A.1.1 Screen update rate

Figure A.1 shows the coding that gather performance data during a normal tryout of the multimodal application. The first line checks whether a requested picture needed for the GUI is stored locally. If not, a timer is started and then a method for downloading the required picture remotely from a web-server is started. When the method has finished and the picture loaded in to the GUI, the timer is stopped and a time difference is calculated. Based on the required time and the size of the picture measured as bytes stored in the readBytes variable, statistical data is recorded and stored to a file. The method can not be characterized to be scientific, but it gives an rough view on what performance to expect during regular testing of the system.



```
readBytes =GetImageFromFile2Buffer((LPSTR)szBuffer, dwBufferMax, pathfile);

        if(readBytes ==0){

        DWORD dwTimeElapsed;
        DWORD dwOldTime;

        dwOldTime = GetTickCount();

        readBytes =GetImageFromURL2Buffer((LPSTR)szBuffer,dwBufferMax, ihandler};

                dwTimeElapsed = GetTickCount() - dwOldTime;

                DWORD dwTransferSpeed = (8*read_bytes)/(dwTimeElapsed/1000);

                FILE *stream;

                if( (stream = fopen( "\\Temp\\time.txt", "a" )) != NULL )
                        {
                                fprintf(stream, "\%s ", fil);
                                fprintf(stream, "\%s ", " size: ");
                                fprintf( stream, "\%d ", read_bytes );
                                fprintf(stream, "\%s ", " time used ");
                                fprintf( stream, "\%d ", dwTimeElapsed );
                                fprintf(stream, "\%s", " ms ");
                                fprintf(stream, "\%d ", dwTransferSpeed);
                                fprintf(stream, "\%s\n", " bits/s ");
                                fclose( stream );
                                        }

                        }
```

Figure A.1: Source code for measuring display update rate

Figure A.2 is a bar chart illustrating the data gathered for both the combined (indicated with CS) and the all-IP (indicated with PS) solution. The figure helps to illustrate that in general, the PS solution provides faster update of graphics. A more drastic result one can extract is that it can take as much as 11.5 seconds to update the screen in worst case.

Based on the size of the pictures download, a throughput rate is also calculated. Natural, as figure A.3 shows, the CS graph provides a lower throughput rate than the PS graph. It can also be seen that the throughput rate varies from just a 5 kbit/s to 50 kbit/s for CS and around 10 kbit/s to almost 100 kbit/s for PS. These results can also depend on the size of the pictures, i.e. small pictures get a relative low throughput rate compared to larger size pictures because some small amount of time is used to start the picture download.

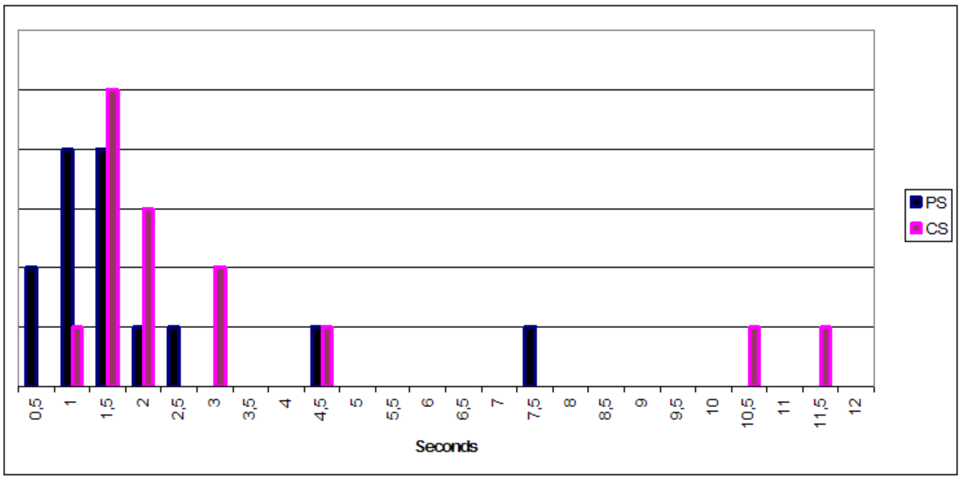


Figure A.2: Frequency of different display update rates divided in intervals of half a second.

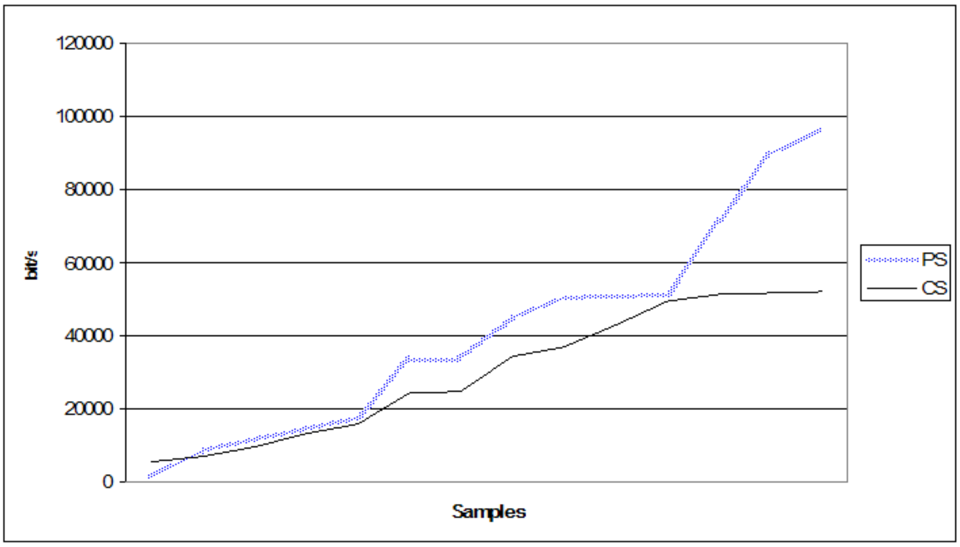


Figure A.3: Obtained data rate during testing measured in kbit/s

A.2 Network measurement

To get a understanding of different aspect of the communication between client and server in the multimodal system, it is important to have a understanding of the underlying network layer providing connectivity service. It has been performed network measurements to view different practical characteristics of the mobile network that may have influence on a multimodal session.

A.2.1 Delay

The delay is an important factor in a simultaneous coordinated multimodal interface and especially for real-time traffic. The measuring of delay between the client and server was done using a basic PING[1] analysis. The PING tool measure the round-trip(RTT), so the one way delay will be the RTT divided by two.

The test bench was set up to resemble the multimodal system and is illustrated in figure A.4. The PING request is sent to a web server connected to Telenor's corporate network. This is done to resemble the path the communication between client and server in the multimodal system follows, because the multimodal server is installed on the same network.

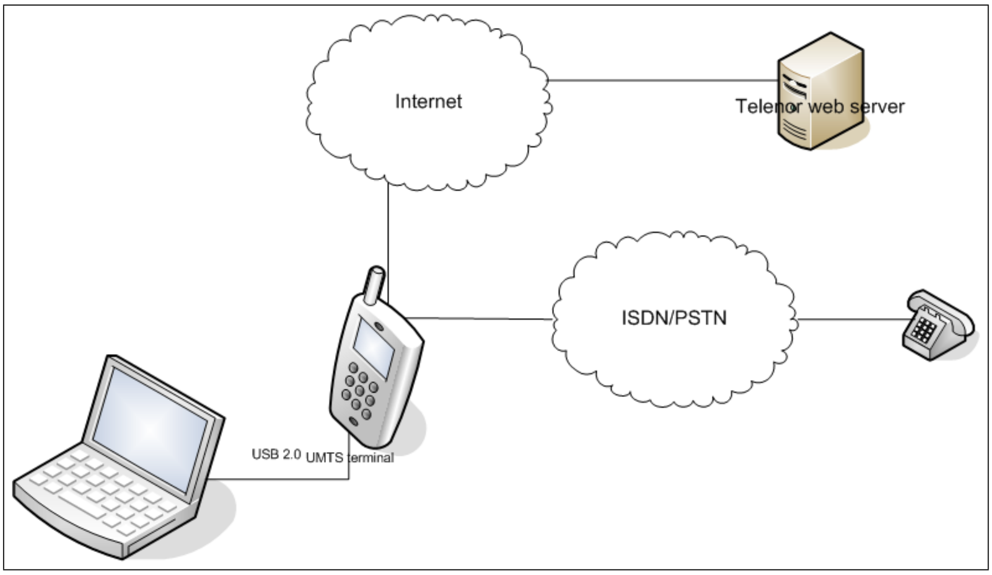
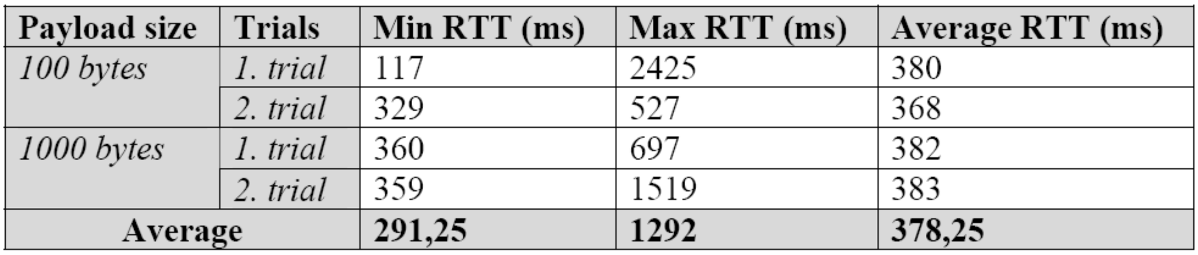


Figure A.4: A coarse network setup overview for testing RTT.

The objective of the test was the measure the delay between two end-nodes over a packet-switched UMTS data connection. The test was performed in two stages. Packets and data transferred between client and server comprise different sizes depending on the application. Some packet may be quite small such as voice-traffic which needs to be sent quite often. Packets containing graphics are more likely to be larger in size. Hence the test has been done in two stages for both connection setup. First test 100 packets with a payload of 100 bytes was sent. Then 100 packets with a payload of 1000 bytes was sent. The test was performed with two different connection settings as described below.

The first test was done when the terminal was connected with only a UMTS packet-switched connection. And table A.1 shows the result of both test and the average of the two tests.

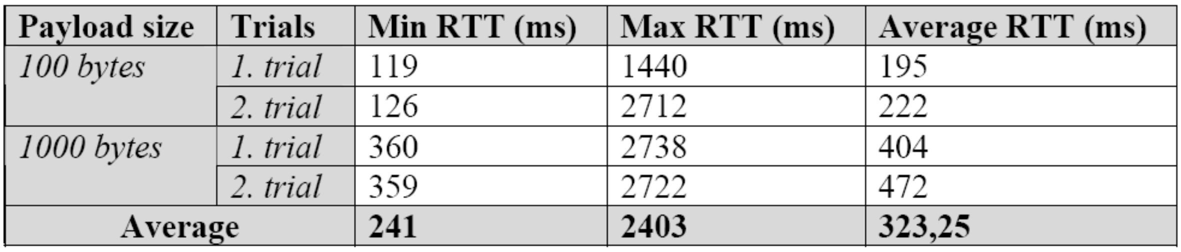
[1] PING is a computer network tool that estimates the round-trip time and packet loss rate between hosts.



Payload size	Trials	Min RTT (ms)	Max RTT (ms)	Average RTT (ms)
100 bytes	*1. trial*	117	2425	380
	2. trial	329	527	368
1000 bytes	*1. trial*	360	697	382
	2. trial	359	1519	383
Average		**291,25**	**1292**	**378,25**

Table A.1: RTT delay for a packet-switched UMTS connection

The second test was done when the terminal was simultaneous connected over a circuit-switched voice connection, i.e. performing a telephone call. The table A.2 show the result of the test.



Payload size	Trials	Min RTT (ms)	Max RTT (ms)	Average RTT (ms)
100 bytes	*1. trial*	119	1440	195
	2. trial	126	2712	222
1000 bytes	*1. trial*	360	2738	404
	2. trial	359	2722	472
Average		**241**	**2403**	**323,25**

Table A.2: RTT delay for a packet-switched UMTS connection when the UMTS terminal is performing telephony

A.2.2 Throughput rate

The throughput rate of the packet-switched UMTS data connection is specified in the UMTS standard. But it is interesting to measure the realistic throughput rate. The throughput available for the packet-switched connection through a 3G terminal was tested using a speed-test tool on a Norwegian computer-oriented website called ‘www.dinside.no’ [9]. Suchlike test tools can not be considered as 100% reliable. But the test give an indiacation of what throughput rate to expect. The test configuration is illustrated in figure A.5.

The UMTS terminal is set up as a modem and connected to a computer via an USB 2.0 connection. The tests were performed inside a building in a densely populated area just outside Oslo. The signal strength indicator showed full strength. The throughput was first tested when simultaneously connected with a circuit-switched voice-connection. The test results is quoted in figure A.6.

Based on doing several tests, the chart shows a average result from the test. The first column show the downlink and uplink throughput when the UMTS terminal is just connected with a

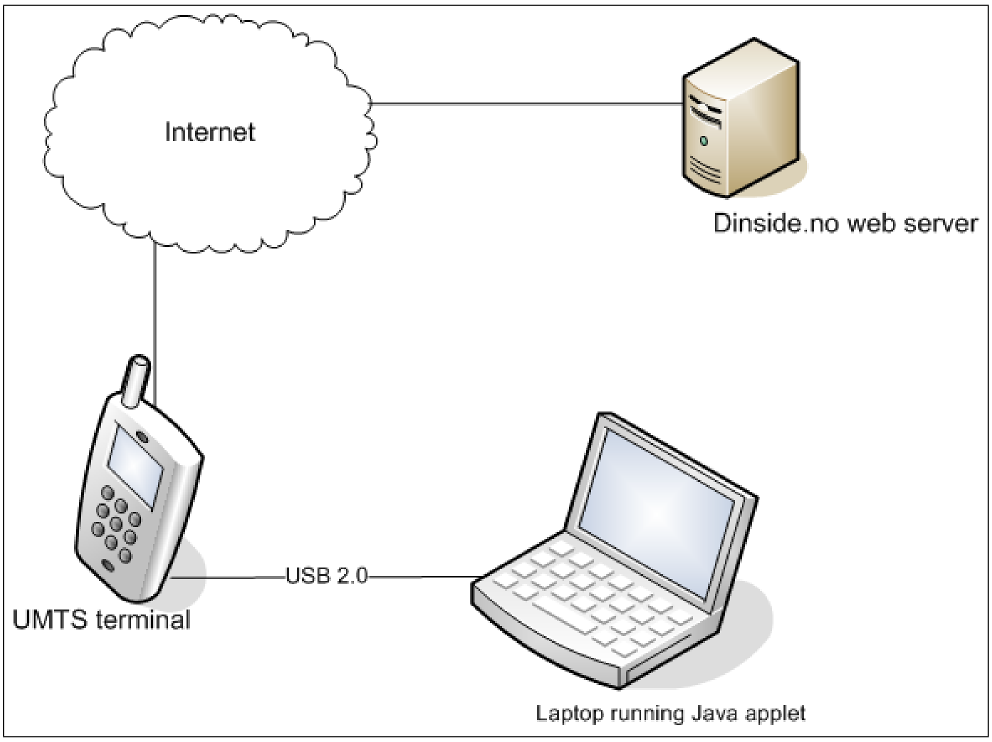


Figure A.5: A coarse network setup overview for testing throughput rate.

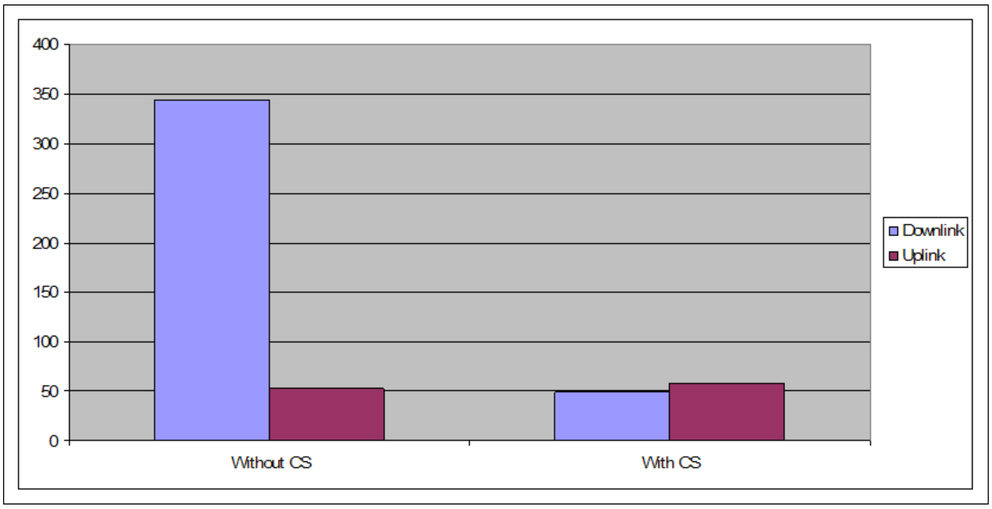


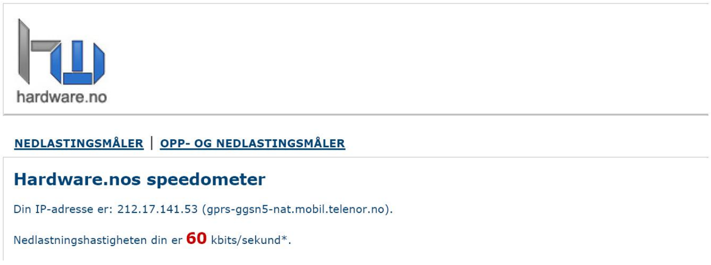
Figure A.6: Average throughput rate for a UMTS data connection

packet-switched UMTS data connection. The second column shows the results when the terminal is connected with both a packet-switched connection and a circuit-switched connection. The first test shows result as expected. Even though UMTS release 99 specifies a downlink rate of 384 kbit/s, a rate just below 350 kbit/s not bad considering the test is performed via a wireless link. The uplink rate is just over 50 kbit/s compared to a maximum throughput according to the specifications of 64 kbit/s. Hence, the results is nearly what one could have expected.

For the second test when the terminal is simultaneously connected to a arbitrary choosen circuit switched telephone, the data rate results are somewhat surprising. The downlink rate is just below 50 kbit/s and the uplink is just above the 50 kbit/s limit. The reason for the downlink rate is lower than the uplink rate may be somewhat arbitrary and many different physical aspects may occur.

The drop in the downlink rate for a simultaneous circuit- and packet-switched is quite drastic. The reason for this result is not 100% clarified. An investingation to resolve this outcome have resultet in different answers. Some experts have said that it is the network management that automatically reduces the throughput rate available for the termnianl. Other experts say that the reduction in data rate is performed by the UMTS terminal for different reasons, e.g. to save battery capacity and because the processing power in the terminal is insufficient for this terminal modus. Nevertheless, the important lesson to learn is that the throughout rate is decreased and that this fact needs to be considered when delivering services operating in similar settings.

To make sure that the results generated from dinside.no was correct, a similar test was performed on an equal tool. Figure A.7 shows the downlink rate for a when the UMTS terminal is connected to both circuit- and packet-switched connection. Figure A.8 shows the downlink rate for the UMTS packet-switched connection when only connected over that connection. The result from these tests confirms the previous test. Figure A.7 gives downlink rate of 60 kbit/s, which indicates that the maximum rate for such setup is restricted to 64 kbit/s under ideal conditions.

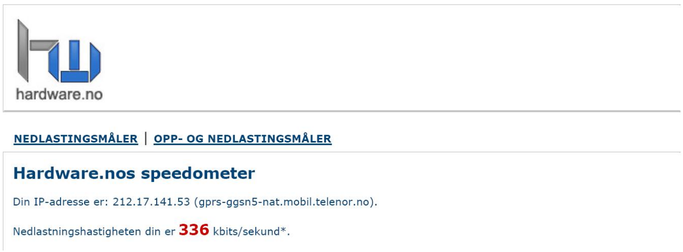


Hardware.nos speedometer

Din IP-adresse er: 212.17.141.53 (gprs-ggsn5-nat.mobil.telenor.no).

Nedlastningshastigheten din er **60** kbits/sekund*.

Figure A.7: Result from testing the UMTS data connection with a simultaneous circuit-switched connection.



Hardware.nos speedometer

Din IP-adresse er: 212.17.141.53 (gprs-ggsn5-nat.mobil.telenor.no).

Nedlastningshastigheten din er **336** kbits/sekund*.

Figure A.8: Result from testing the UMTS data connection without a simultaneous circuit-switched connection.

Appendix B

Manual for the MustClient

This section provides installations steps to try out the multimodal Bus Demo service. This requires that you have a UMTS terminal running Microsoft Windows Mobile 5.0 Pocket PC edition.

B.1 Install the MultimodalClient

Follow this procedure to install the multimodal client.

1. Connect the terminal to a PC running Windows and ActiveSync installed.
2. Copy the file ‘MultimodalClient.cab’ from the folder ‘multimodal’ contained in the ZIP-file to a temporary folder on the terminal.
3. Disconnect the terminal from the PC.
4. Open the ‘MultimodalClient.cab’ file you transferred.
5. The installation of the MustClient will begin. When finished, choose OK.
6. Go to the folder ‘Program Files’ and then ‘MultimodalClient’.
7. Start the application by pressing the ‘MustClient.exe’ file.

B.2 Remove the MultimodalClient

To remove the application, follow this procedure:

1. Go to ‘Settings’ under the ‘Start’ button on your terminal.
2. Choose the ‘System’ tab and open the ‘Remove Program’.
3. Locate ‘Telenor R&D MultimodalClient’ in the Program list.
4. Highlight the program and click ‘Remove’.
5. Confirm ‘Yes’ to remove the program and the application is removed from your terminal.

B.3 How to use the Bus demo application

This section will describe how the demonstrator application ‘Bus Demo’ can be operated. In this application the user has the possibility to see how a multimodal interaction happens and especially try the coordinated simultaneous multimodal interaction method which utilize the multimodal system to the fullest. If the reader would like to try out the service, please contact the author on ‘thormod.schie@gmail.com’ to schedule a time for trying the service. The server running the service is stationed at Telenor R&D and is used for other demonstrator purposes also, so some setup time is required to prepare the server for a MUST demo.

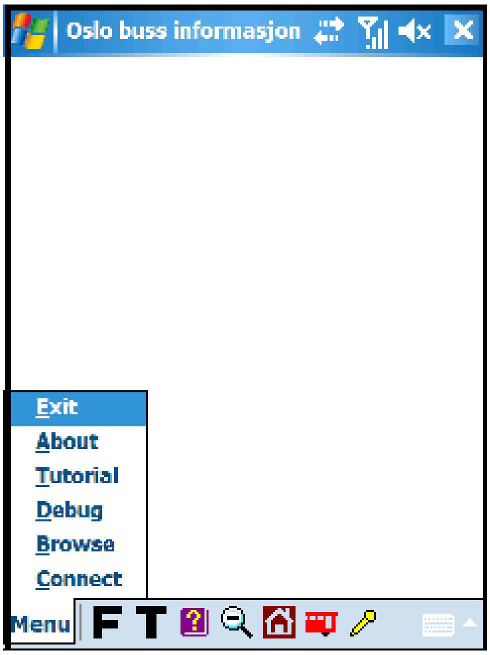


Figure B.1: Application opened

When the user opens the application, the application will present a picture similar to figure B.1. Then the user must click the connect button on the menu to connect to the multimodal server.When the user is logged in to the server the user will be met by screen like figure B.2.Here the user can either do a pure speech interaction with the system asking for example: “When does the next bus goes from Fornebu to Majorstuen?” or he can click on one of the sub-maps indicated with the black lined squares.

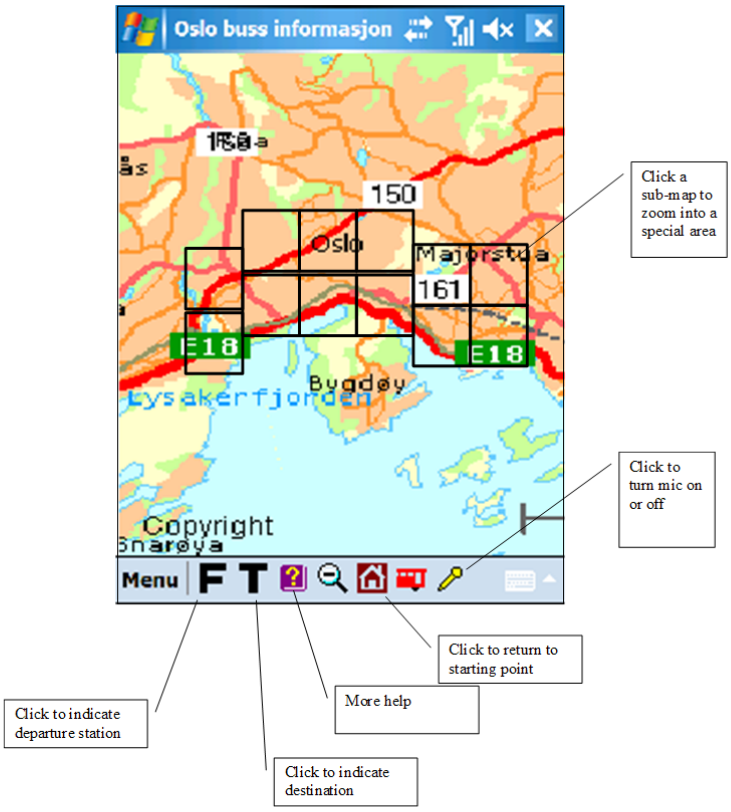

Oslo buss informasjon
150
Oslo
Majorstua
161
E18
E18
Bygdøy
Lysakerfjorden
Copyright
Snarøya
Menu
F
T
Click a sub-map to zoom into a special area
Click to turn mic on or off
Click to return to starting point
More help
Click to indicate departure station
Click to indicate destination

Figure B.2: Frontpage of the application

If the user choose to click on one of the sub-maps, the screen will be updated and display a screen similar to B.3. Here the user can click on one of the bus station icon and simultaneously ask: “When does the next bus leaves from here to Majorstuen?” This will then correspond to a coordinated simultaneous multimodal interaction as described in 2.1.3.3. Or he can click on either the symbol F or T corresponding to ‘From’ and ‘To’ respectively, then a bus station icon to indicate departure or arrival station. And to the same operation for the other desired bus station. This operation involves no speech, hence it corresponds to a pure pen interaction. To zoom out, the user can touch the zoom out button or simply say: “Zoom out”. Every commands available by using the pen is also available with speech instructions.

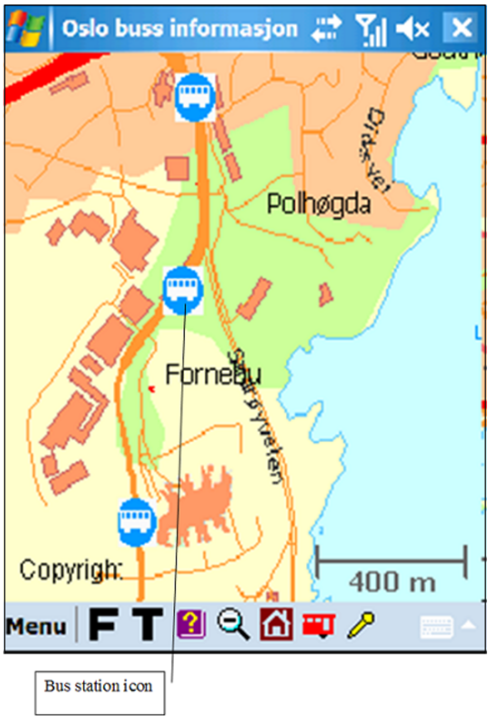


Figure B.3: A sub-map marked with bus stations

Appendix C

Enclosed ZIP-file

The ZIP-file is divided in following folders containing:

- src - Source files for the client software.
- stat - Text files containing results from network measurement.
- reference - References used in the report that are stored in an electronic format.
- program - Folder containing installation file for ‘MultimodalClient’.
- report - Containing an electronic version of this report.
- video - A film clip demonstrating the multimodal application.