13

Multimodal Interfaces to Mobile Terminals – A Design-For-All Approach

Knut Kvale and Narada Dilp Warakagoda
Telenor Group Business Development & Research
Norway

1. Introduction

Multimodal human-computer user interfaces are able to combine different input signals, extract the combined meaning from them, find requested information and present the response in the most appropriate format. Hence, a multimodal human-computer interface offers the users an opportunity to choose the most natural interaction pattern for the actual application and context of use. If the preferred mode fails in a certain context or task, users may switch to a more appropriate mode or they can combine modalities.

Around thirty years ago Bolt presented the "Put That There" concept demonstrator, which processed speech in parallel with manual pointing during object manipulation (Bolt, 1980). Since then major advances have been made in automatic speech recognition (ASR) algorithms and natural language processing (NLP), in handwriting and gesture recognition, as well as in speed, processing power and memory capacity of computers. Today's multimodal systems are capable of recognizing and combining a wide variety of signals such as speech, touch, manual gestures, gaze tracking, facial expressions, head and body movements. The response can be presented by e.g. facial animation in the form of human-like presentation agents on the screen in a multimedia system. These advanced systems need various sensors and a lot of processing power and memory. They are therefore best suited for interaction with computers and in kiosk applications, as demonstrated in e.g. (Oviatt, 2000); (Gustafson et al., 2000); (Wahlster, 2001); (Beskow, et al. 2002); (Karpov, 2006); (Smartkom, 2007).

Modern mobile terminals are now portable computers where the traditional audio user interfaces, microphones and speakers, are accompanied with touch screens, cameras, accelerometers and gyroscopes etc. These enriched user interfaces combined with the ever increasing capacity of processors, access to mobile networks with increasing bandwidths and functionality as global positioning system (GPS) and near field communication (NFC) will make mobile terminals well suited for developing user-friendly multimodal interfaces in the years to come.

However, the multimodal functionality on mobile terminals is still restricted to two input modes: speech (audio) and touch, and two output modes: audio and vision. This type of multimodality, sometimes called tap & talk (or point & speak), is essentially *speech centric*, and will be explored further in this chapter.

We will investigate the hypothesis that multimodal interfaces offer a freedom of choice in interaction pattern for all users. For normal able-bodied users this implies enhanced user-

Source: User Interfaces, Book edited by: Rita Mátrai,
ISBN 978-953-307-084-1, pp. 270, May 2010, INTECH, Croatia, downloaded from SCIYO.COM

friendliness and flexibility in the use of the services, whereas for the disabled users this is a means by which they can compensate for their impaired communication mode.
The outline of this chapter is as follows: Section 2 first defines multimodal interaction and discusses various forms of multimodality. Then we confine ourselves to speech centric multimodal interfaces for mobile terminals and demonstrate the advantages of this functionality in two form-fillings applications. Section 3 relates the principles of Design for All to multimodal user interfaces. Section 4 presents a generic system architecture for multimodal interfaces, whereas Section 5 provides more details of our implementation of a public web-based bus-route information service. Section 6 describes the user evaluations of our system by five test persons with different impairments, as well as a dyslectic and an aphasic test user.

2. Various forms of multimodality

2.1 Multimodal versus multimedia

The term modality refers to a form of sensory perception: hearing, vision, touch, taste and smell. For our research on human-machine interaction, we define modality as a communication channel between the user and the device. The modes above can be combined in a multimodal interface, containing audio (e.g. in the form of speech), vision (e.g. in the form of text and graphics, or moving video), and touch (e.g. touch sensitive screens). We do not consider services using one particular input mode, e.g. speech, and another output mode, e.g. text/graphics as multimodal services. We distinguish between multimode and multimedia; that is, media is the representation format for the information or content in a certain mode. For example, speech and music are two media formats in the auditory mode. Text, graphics and video are examples of media types in the visual mode.

2.2 Combining multiple modalities

Multiple input and output modalities can be combined in several ways. We may distinguish between combining the multimodal inputs sequentially or simultaneously. In a sequential multimodal system inputs from different modalities are interpreted separately. For each dialogue state, there is only one input mode available, but in the whole interaction more than one input mode may be used. Sequential multimodal input is often used in system-driven applications. Some systems may offer several parallel input modes that are active at the same time. This means that the users can choose the input mode they prefer at each dialogue stage. However, only one of the input channels is interpreted (e.g. the first input).
In a simultaneous multimodal system, also called composite multimodality, all inputs within a given time window are interpreted jointly depending on the fusion of the partial information from the different input channels. Composite multimodality is probably the most natural way of interacting with computers, but it is by far the most complicated scenario to implement.
On the output side the difference between sequential and simultaneous use of modes may be less apparent, because the graphical display is static: it remains visible during times when speech is played (and the graphical image cannot be changed). In coordinated simultaneous multimodal output, information may be conferred by means of a spoken message that coincides with changes in the graphical display and perhaps also with gestures of an on-screen presentation agent.

2.3 Mobile terminals and multimodality

The first generation of small mobile terminals used for mobile communication purposes had only a handful of input and output modalities: e.g.: speech and a small key pad on the input side and a small black and white character display and audio on the output side. The simplicity of the task they were meant to be used for, namely to make or answer a call had probably justified such a very simple user interface. But the tasks of the mobile terminals quickly started to get more complex and the need for more sophisticated user interfaces started to grow. This need has been addressed to a certain extent by the technological development in the past decade or so, even though the user interfaces could never keep up with the development of the functionalities of the mobile terminals.

One significant development in user interfaces of the mobile terminal is its screen. Presently almost all small mobile devices are equipped with high resolution colour screens capable of rendering advanced graphics. While this is a huge boost of the user interface on the output side, another property of the screens, namely touch sensitivity, contributed heavily to improving the input side. In 2002/2003 high end mobile terminals with touch sensitive screens appeared in the market (e.g. Sony Ericsson P800 (GSM, 2009)). But now the touch screen is a common feature of even mid-range mobile devices.

In the latter part of this decade, several other user interface components integrated in mobile terminals became very common. One such component is the camera. This can provide the basis for implementation of input modalities such as object recognition, face recognition and gaze tracking etc.

Another common integrated component in modern terminals is the Global Positioning System (GPS) receiver module, which can provide the location information, essentially an input modality. So-called Near Field Communication (NFC) technology which is expected to be a common feature of mobile terminals in the next two to three years is another way of getting location information. NFC is often considered to be a technology supporting the pointing modality and can be used in novel multimodal applications as voice-enabled mobile commerce (Warakagoda et al., 2008).

Even though most of the above mentioned user interface technologies have existed in a sufficiently mature state for a fairly long time, there hasn't been any breakthrough in the user interfaces of mobile terminals until Apple's iPhone was introduced in 2007 (GSM, 2009). Worldwide success of this product was mainly due to its attractive user interface combining several technologies mentioned above. The iPhone exploits the touch sensitive screen in a clever way, not only to support pointing but also touch gestures. In addition, the iPhone makes use of microelectromechanical systems (MEMS) technology such as accelerometers and gyroscopes to create completely new modalities like acceleration and orientation.

Inspired by the success of iPhone, a wave of similar devices has been released into the market by the rivalling manufacturers. The result is that now we have a large number of mobile device models which include user interface modules such as touch screens, GPS, cameras, accelerometers and gyroscopes etc. We should not forget that the traditional audio devices, microphones and speakers are still there and the mass market NFC is just around the corner. On top of all these, the modern mobile devices are equipped with high capacity processors and network interfaces such as Universal Mobile Telephony System (UMTS) and High Speed Packet Data Access (HSPA). All those factors make today's mobile terminals an ideal platform for developing multimodal interfaces.

2.4 Speech centric multimodality

The full potential of all the functionality described in section 2.3 above is not exploited yet. The multimodal functionality on mobile terminals is still usually restricted to two input modes: speech (audio) and touch, and two output modes: audio and vision. This type of multimodality, sometimes called tap & talk (or point & speak), is essentially *speech centric*, and will be explored further in this chapter.

In most speech centric multimodal interfaces on mobile terminals, the input combines and interprets spoken utterances and pen gestures such as pointing, circling and strokes on a touch sensitive screen. The output information is either speech (synthetic or pre-recorded) or text and graphics.

Speech centric multimodality utilises the fact that the pen/screen and speech are *complementary*. The advantage of pen input and screen output is typically the weakness of speech, and vice versa: Spoken interaction is temporal, whereas visual interaction is spatial. With speech it is natural to ask one question containing several key words, but it may be tedious to listen to all information read aloud because speech is inherently sequential. With pen and graphics interfaces only, it may be hard to enter queries, but it is easy to get a quick overview of the information on the screen, as summarised in Table 1.

Only pen input, screen output	Pure speech input/output
Hands and eyes busy – difficult to perform other tasks	Hands and eyes free to perform other tasks
Simple actions	Complex actions
Visual feedback – spatial	Oral feedback – temporal
No reference ambiguity	Reference ambiguity
Refers only to items on screen	Natural to refer also to invisible items
No problem with background noise	Recognition rate degrades in noisy environments

Table 1. Comparison of the two complementary user interfaces: Pen only input and screen output versus a pure speech based input and output interface.

Hence, systems combining the pen and speech (tap & talk) input may lead to a more efficient human-computer dialogue:

- The users can express their intentions using fewer words and selecting the input mode they judge to be less prone to error, or switch modes after system errors and thus facilitate error recovery.
- The system offers better error avoidance, error correction and error recovery.

Speech centric multimodal interfaces for mobile terminals can be utilised in many different applications. In e.g. (Watanabe et al., 2007), the complementary merits of speech and pen are utilised for entering long sentences into mobile terminals. With this interface, a user speaks while writing, where the two modes complement one another to improve the recognition performance. However, the two most promising mobile applications with speech centric multimodality are *form-filling* and *map-based systems*.

2.5 Speech centric multimodality for form-filling

In this section we exemplify the benefits of speech centric multimodality in two form-filling applications on a wireless personal digital assistant (PDA) with touch sensitive screen: A public train timetable information retrieval service and a public "yellow pages" service.

Figure 1 below shows the graphical user interface (GUI) in three dialogue steps of the service for a Norwegian train timetable information retrieval application:

1. This entry page appears on the screen when the service is called up. Below the text heading: "Where do you want to go?" there are five input form fields: Arrival and departure station, date and time of arrival and the number of tickets. The questions are also read aloud by text-to-speech synthesis (TTS).
2. This screen shows the result of the user request in natural language: "I want to go from Kristiansand to Bodø next Friday at seven o'clock". The key words in the utterance were recognised correctly and the corresponding fields filled in, giving the user an immediate feedback on the screen. The call was made on June 10, so "next Friday" was correctly interpreted as June 15.
   Since all the information in the form fields on the screen is correct the user confirms by pushing the 'OK' button, and the system gets the requested information from the railway company web portal.
3. The result of the web request is presented on the screen. Usually three or four realistic alternatives are depicted on the screen. The user may then tap on the preferred travel alternative, or say the alternative number. Then the dialogue goes on asking "how many tickets" the customer wants for the selected trip and this demonstrator service ends.

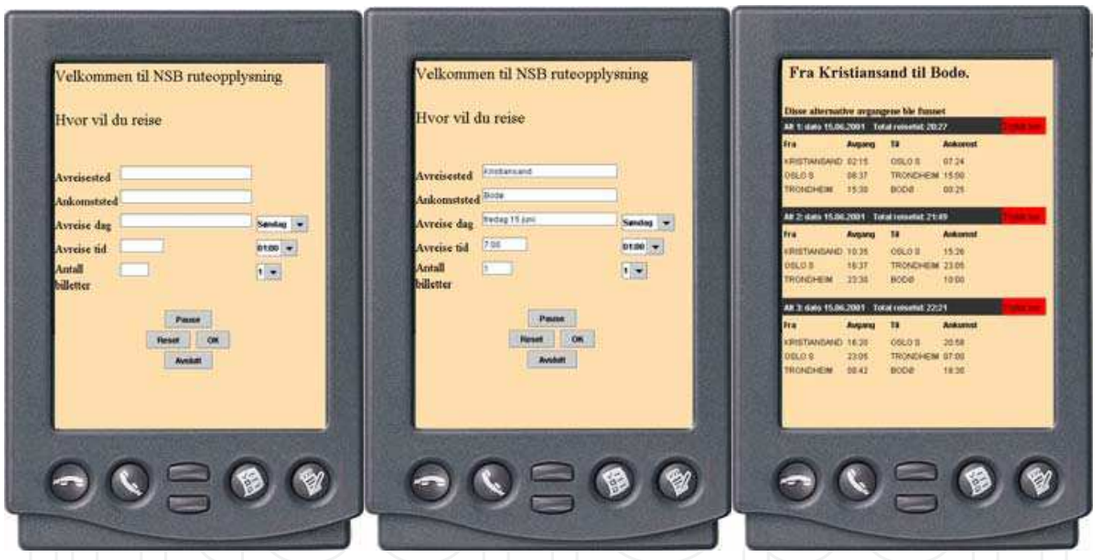

Fig. 1. The GUI for the train timetable information retrieval application

In the example in figure 1, all the words were correctly recognised and understood and the visual presentation of information was much more efficient than audio feedback. Therefore the customer obtained efficiently what she wanted. However, in real world speech-enabled telephony applications ASR-errors will unavoidably occur. Correcting ASR-errors in speech only mode (no visual feedback) is very difficult and reduces the user satisfaction. But, with a speech centric multimodal interface it is rather easy to correct ASR-errors in these form-filling services. If some of the information on the screen is wrong, the user corrects it by clicking on the field with wrong words and then either saying the correct word once more or tapping on the correct word from the N-best list, which occurs on the right hand side of the field.

Figure 2 illustrates this situation in the "yellow pages" application:

1. The entry page that appears on the screen when the service is called up:
   Below the text heading: "Welcome to Yellow pages" there are two input form fields: Business sector and municipal (Norwegian: "Bransje" and "sted")
2. When the user has asked in natural language: "I want bakeries in Oslo". The ASR recognised the key words in the utterance and filled in the corresponding fields, giving the user an immediate feedback on the screen. Note that the N-best list on the right hand side of the sector field contains the alternative "Batteries". That is, the word "batteries" has the second best confidence score.
   Since all the information in the form fields on the screen is correct the user pushes the 'OK'-button, and the system gets the requested information from the service provider.
3. The requested information is displayed on the screen. There are 25 bakeries in this listing which would have been rather tedious listening to. Here, the user easily gets a quick overview and clicks on the preferred baker.

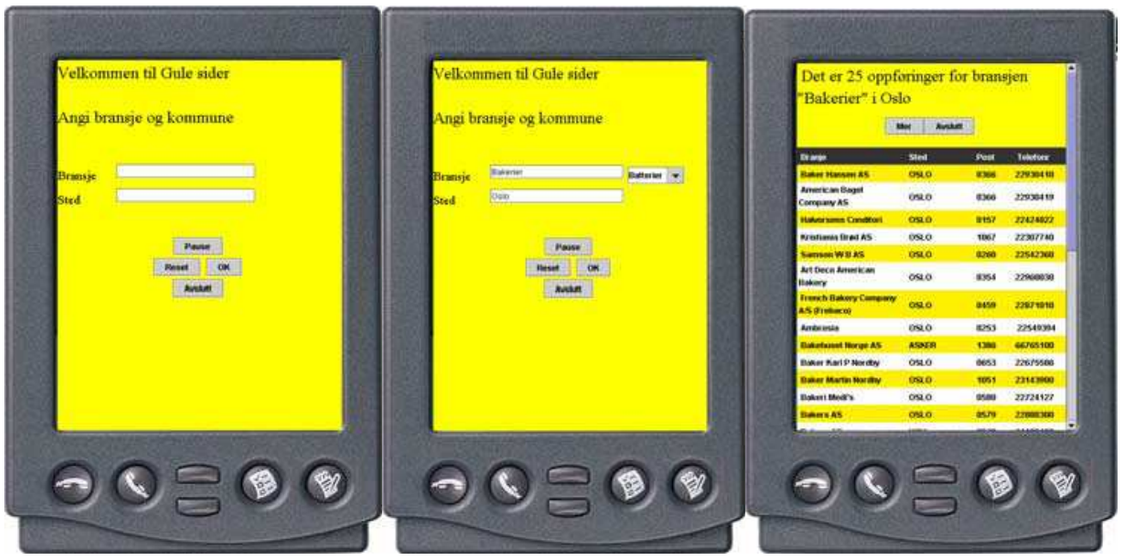

Fig. 2. The GUI for the Yellow pages application

The actions and benefits of speech centric multimodality in the form-filling applications are summarized in table 2.

User actions	Benefits of multimodality
Natural language input, asking for several different pieces of information in one sentence.	Speech is most natural for asking this type of questions. Speech is much faster than typing and faster than selecting in a hierarchical menu.
Reads information shown on the screen.	The user gets a quick overview – much quicker than with aural feedback reading sentence by sentence.
Taps in the field where the ASR-error occur, and taps at the correct alternative in the N-best list.	Much easier to correct ASR-errors or understanding rejections than with complicated speech-only dialogues. Better error control and disambiguation strategies (e.g. when there are multiple matching listings for the user query).

Table 2. Benefits of speech centric multimodality in form-filling applications.

2.6 Speech centric multimodality for map-based applications

Combining speech and pen gestures as inputs to mobile terminals has proven particularly useful for navigation in maps. Typically, this kind of speech centric multimodal mobile applications provides easy access to useful city information, for instance restaurant and subway information for New York City (Johnston et al., 2001), (Johnston et al., 2002), a tourist guide for Paris (Almeida et al., 2002a), (Almeida et al., 2002b), (Kvale et al., 2003b), bus information system for the Oslo area (Kvale et al., 2004), (Kvale et al., 2005), (Warakagoda et al., 2003) navigational inquiries in the Beijing area (Hui & Meng, 2006), trip planning and guidance while walking or driving car (Bühler & Minker, 2005), various maptasks with "QuickSet" (Oviatt, 2000) and services aimed at public transportation commuters (Hurtig, 2006). Task analysis of map interfaces have shown that multimodal interaction with tap and talk is natural during spatial location and selection commands such as: "What's the distance from here to here?" <while tapping at actual locations in the map>, or "Zoom in this area" <while tapping at the area on the map>.
Our bus information system for the Oslo area fits into these kinds of applications and will be discussed further in Section 5 and 6.

3. Multimodal interfaces are useful for all

Tim Berners-Lee, one of the inventors of the World Wide Web, stated in 1997 that "The power of the Web is in its universality. Access by everyone regardless of disability is an essential aspect". However, accessibility to web based information services is still limited for many people with sensory impairments. A main obstacle is that the input and output channels of the services support one modality only. The European Telecommunications Standards Institute has claimed that the missing access to environments, services and adequate training contributes more to the social exclusion of disabled people than their living in institutions (ETSI, 2003).
There are two different approaches to solving this problem. One is to develop special assistive technology devices that compensate for or relieve the different disabilities. Another solution is to design services and products to be usable by everybody, to the greatest extent possible, *without* the need for specialized adaptation; a so-called Design-for-All approach.
Design for All (DfA), also called Universal Design, is a user-centred design approach which addresses the possible range of human abilities, skills, requirements, and preferences. There exist a lot of guidelines and principles for DfA, as for instance the seven principles for universal design proposed by the Centre for Universal Design North Carolina State University (NC, 1997), and the Web Content Accessibility Guidelines (WCAG) developed by W3C (WCAG, 2008). Following these guidelines will not only make Web content more accessible to a wider range of people with disabilities, it will also often make the Web content more usable and provide all users with a better user experience.
The core of these guidelines and recommendations is to accommodate a wide range of individual preferences and abilities by offering alternative interaction modes and redundancy in the presentations. In our opinion, multimodal human-computer user interfaces have the potential to fulfil the requirements for universal design. Multimodal interfaces are able to combine different input signals, extract the combined meaning from them, find requested information and present the response in the most appropriate format. Hence, a multimodal human-computer interface offers the users an opportunity to choose the most natural interaction pattern depending on the actual task to be accomplished, the

context, and their own preferences and abilities. If the preferred mode fails in a certain context or task, users may switch to a more appropriate mode or they can combine modalities.

We believe that multimodal interfaces offer a freedom of choice of interaction pattern which is useful for all users. For able-bodied users this implies enhanced user-friendliness and flexibility in the use of the services, see e.g. (Kvale et al. (2003b), (Oviatt et al., 2004), whereas for the disabled users this is a means by which they can compensate for their not-well-functioning communication mode.

To test the hypothesis that multimodal inputs and outputs really are useful for disabled people, we have developed a flexible speech centric multimodal interface on mobile terminals to a public web-based bus-route information service for the Oslo area.

4. General implementation aspects of multimodal systems

Figure 3 shows a typical multimodal system architecture. This is essentially an input-output system where multiple inputs are integrated and the result is used to determine the outputs. Inputs can be integrated either before or after they are recognized and semantics are extracted. The former and latter cases are known as *early fusion* and *late fusion* respectively. The dialogue manager functional module generates the response using this fused input and the current context. The response planner module determines how the response is presented to the user by splitting up the semantic stream coming out of the dialogue manager into appropriate modalities. This process is also known as *fission*. Both multimodal integration and response planner typically make use of context information to control their actions. In the following sub-sections the functionalities of the most important modules in figure 3 are explained.

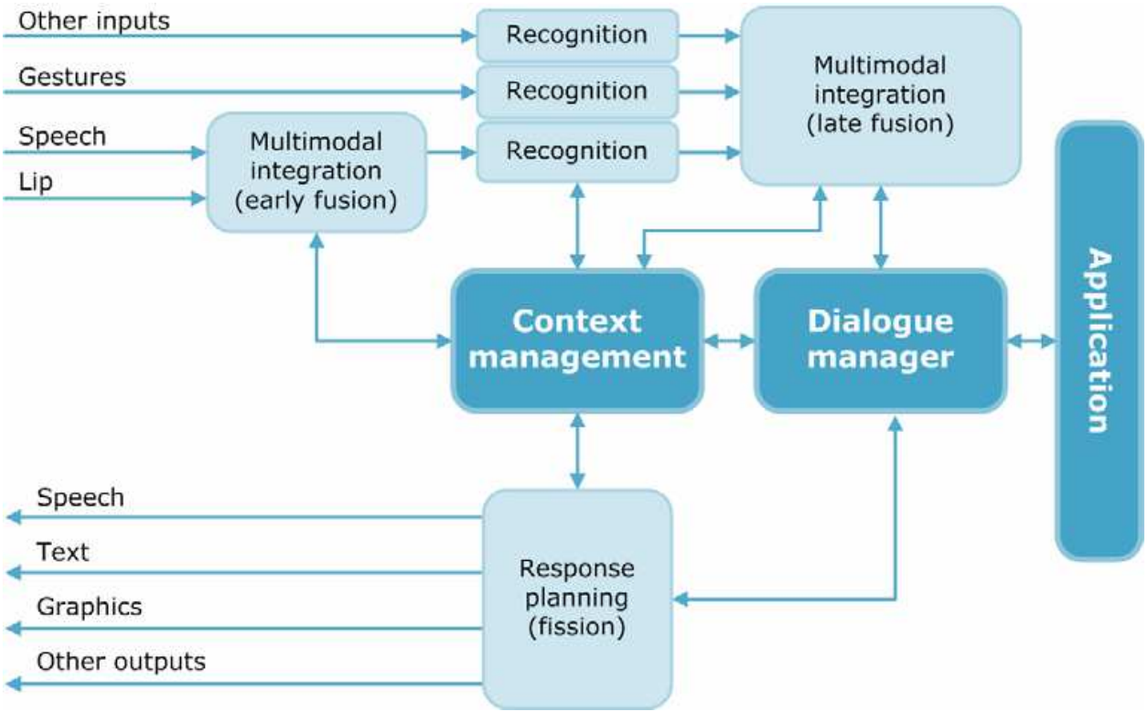

Fig. 3. A generic multimodal dialogue system architecture

4.1 Recognition

This is one of the most important operations in multimodal systems. In recognition an input data stream is classified into a predefined number of classes and the resulting class labels are mapped on to a vector of semantic units. The position of each element of this vector corresponds to a semantic concept and the element itself is the value of the corresponding semantic concept. Often, recognition is a statistically based process and therefore the outcome of recognition is not a single concept vector, but a list of vectors where each of these vectors is associated with a probabilistic score or likelihood. For example, suppose that the input is a speech signal and the recognizer is designed to recognise utterances such as "I would like to take a bus from Oslo to Fornebu 10 o'clock today". Then a suitable set of semantic concepts would be (<FROM_PLACE>, <TO_PLACE>, <DEPARTURE_TIME>). If the user has actually uttered the above sentence, and we limit the output to three concept vectors, examples of output can be:

- (Oslo, Fornebu, 1000) with probabilistic score 0.18
- (Oslo, Fornebu, 1300) with probabilistic score 0.11
- (Oslo, Fornbuveien, 1000) with probabilistic score 0.09

4.2 Fusion and fission

Since a multimodal system has more than one input and/or output channel, there must be mechanisms to map:

- Several input channels to a single semantic stream, i.e. fusion
- Single semantic stream to several output channels, i.e. fission.

From a technical point of view, fusion, also called multimodal integration, deserves a higher attention than fission, because a good fusion strategy can help reduce the recognition errors. Usually, fusion is classified into two classes, early fusion and late fusion. Early fusion means integration of the input channels at an early stage of processing. Often, this means integration of feature vectors before they are sent through the recogniser(s). Late fusion means integration of the recogniser outputs, usually at a semantic interpretation level. Late fusion seems to have attracted more interest than early fusion, probably because it only needs the recogniser outputs, and no changes of existing modules (such as feature extractors, recognisers) are required.

In one of its simplest forms, late fusion can be performed by simple table look-ups. For example, assume that we have two input channels. Then we can maintain a two dimensional table, whose rows and columns correspond to alternative outcomes of the recognisers acting on channel 1 and channel 2 respectively. Each cell of the table can be marked 1 or 0, indicating whether this particular corresponding combination is valid or invalid. Then the fusion procedure for a given pair of recogniser output lists would be to scan the (recogniser) output combinations in decreasing order of likelihood or probabilistic score and find the first valid combination by consulting the table.

In the above procedure, likelihood is derived from the joint probability of the recogniser outputs from the two channels. One simple approach of computing these joint probabilities is to assume that two recognition streams are statistically independent. However, the fusion performance (i.e. multimodal recognition performance) can be enhanced by dropping this assumption in favour of more realistic assumptions (Wu et al., 1999).

Table look-up based fusion is not very convenient when the semantic information to be integrated is complicated. In such cases typed feature structures can be used. This data

structure can be considered as an extended, recursive version of attribute-value type data structures, where a value can in turn be a feature structure. Typed feature structures can be used for representing meaning as well as fusion rules. Integration of two or several feature structures can be achieved through a widely studied algorithm called feature-structure unification (Oviatt et al., 2000).

In fusion, temporal relationships between different input channels are also very important. This issue is usually known as synchronization. In most of the systems reported in the literature, synchronization is achieved by considering all input contents that lie within a pre-defined time window. One can do this very easily by employing timers and relying on the real arriving times of the input signals to the module responsible for performing fusion. However, a more accurate synchronization can be obtained by time-stamping all inputs as soon as they are generated since this approach will remove the errors due to transit delays. Note however, that input synchronization is meaningful only for coordinated multimodality.

4.3 Dialogue management

The dialogue manager is usually modelled as a finite state machine (FSM), where a given state S_t represents the current context. One problem of this modelling approach is the potentially large number of states even for a relatively simple application. This can be brought to a fairly controllable level by considering a hierarchical structure. In such a structure there are only a few states at the top level. But each of these states is thought to be consisting of several sub states that lie in the next level. This can go on until the model is powerful enough to describe the application concerned.

When the user generates an event, a state transition can occur in the FSM describing the dialogue. The route of the transition is dependent upon the input. This means that state transition is defined by the tuple (S_t, I_t), where S_t is the current state and I_t is the current user input. Each state transition has a well-defined end state S_{t+1} and an output O_t. In other words, the building-block-operation of the dialogue manager is the following:

1. Wait for input (I_t)
2. Act according to (S_t, I_t), for example by looking-up a database and getting the result (R_t)
3. Generate the output according to (S_t, I_t , R_t)
4. Set next state S_{t+1} according to (S_t, I_t)

The user input (I_t) is a vector which is a representation of the structure called concept table. This structure consists of an array of concepts and the values of each of these concepts. For example in a travel planning dialogue system the concept table can look as follows:

Concept	Value
<FROM_PLACE>	Oslo
<TO_PLACE>	Fornebu
<DEPARTURE_TIME>	1600

The column "value" of the concept table is filled using the values output by the recognisers operating on the input modalities (e.g.: speech and GUI tap recogniser). Late fusion completes the filling operation by resolving input ambiguities and ensuring a concept table of the highest likelihood. Once filled, the concept table defines the current input I_t. More specifically, if the values in the concept table are $I_{t(1)}, I_{t(2)}$, … … $I_{t(n)}$, then the N-tuple $(I_{t(1)}, I_{t(2)}$, … … $I_{t(n)})$ is the current input I_t. The number of different inputs can be prohibitively large,

even if the length of the concept table (M) and the number of values a given concept can take (K) are moderate. This implies that a given state in the dialogue FSM has a large number of possible transitions.
A possible remedy for this problem is to employ a clever many-to-one mapping from the original input space to a new smaller sized input space, which exploits the fact that there are many don't-care concept values.

4.4 Internal information flow

In advanced multimodal systems, several input/output channels can be active simultaneously. In order to cope with this kind of multimodality, an architectural support for simultaneous information flow is necessary. Furthermore, it is desirable to run different functional modules separately (often on different machines), in order to deal with the system's complexity more effectively. The so-called distributed processing paradigm matches these requirements quite nicely, and therefore most of the multimodal system architectures are based on this paradigm.
There are many different approaches to implementing a distributed software system. Examples are Parallel Virtual Machine (PVM), Message Passing Interface (MPI), RPC-XML and SOAP, CORBA, DCOM, JINI and RMI (Kvale et al., 2003a). However, a more attractive approach to implementation of multimodal systems is based on co-operative software agents. They represent a very high level abstraction of distributed processing and offer a very flexible communication interface.
There are several agent architectures that have been used to build multimodal systems, for instance GALAXY Communicator from MITRE (Galaxy, 2007), Open Agent Architecture (OAA) from SRI international (OAA, 2009) and Adaptive Agent Architecture (AAA) from Oregon Graduate Institute (Kumar et al., 2000). In these architectures a set of specialized agents are employed to get different tasks performed. Two given agents can communicate (indirectly) with each other through a special agent called facilitator.
We found that GALAXY Communicator is the most suitable agent-based platform for our purpose. A more detailed description of this is given in section 5. The GALAXY Communicator has a hub-spoke type architecture and allows easier asynchronous and simultaneous message exchange between modules than for example a serial architecture does. One drawback of GALAXY Communicator, however, is its dependency on a single facilitator whose failure will cause a complete system breakdown. In AAA this problem has been addressed by introducing many facilitators.
Another aspect of information flow between different modules is the format in which information is packaged during transition. In Galaxy Communicator an attribute-value type of format is used. The advantage of this approach is that this format is very similar to the concept table format used in multimodal integration and dialogue management. This issue has attracted the attention of standard developing organizations too. Especially, the W3C Multimodal Interaction Working Group which develops specifications to enable access to the Web using multimodal interaction has addressed this issue in their Extensible MultiModal Annotation markup language (EMMA) standard. In 2009 the W3C Recommendation for EMMA was launched (W3C, 2009). EMMA markup language is intended for use by systems that provide semantic interpretations for a variety of inputs, including but not necessarily limited to, speech, natural language text, GUI and ink input. According to W3C it is expected that EMMA will be used primarily as a standard data

interchange format between the components of a multimodal system; in particular, it will normally be automatically generated by recognition/interpretation components to represent the semantics of users' inputs, not directly authored by developers.

5. Our speech centric multimodal system

Our multimodal bus information system implements the functional architecture described in section 4 through a set of software modules. Our implementation consists of a server and a thin client (i.e. the Mobile Terminal) as shown in Figure 4. The client server architecture is based on the Galaxy communicator (Galaxy, 2007). The server side comprises five main autonomous modules which inter-communicate via a central facilitator module (HUB) as shown in figure 4. All the server side modules including the automatic speech recogniser (ASR) and the text to speech synthesizer (TTS) run on a PC, while the client runs on a mobile terminal, in this case a Qtek 9000. The client consists of two main components handling voice and graphical (GUI) modalities. It communicates with the server over an Internet Protocol (IP) network such as wireless local area network (WLAN) based on the IEEE 802.11b protocol, or a 3G/UMTS data network. The server communicates with a public web service called "Trafikanten" through the Internet to get the necessary bus route information (Trafikanten, 2009). The "Trafikanten" service is text based (i.e. unimodal). That is, the users have to write the names of the arrival and departure bus stops to get the route information, which in turn is presented as text. Our multimodal interface at the mobile client converts the web service to a map-based multimodal service supporting speech, graphic/text and pointing modalities as inputs. Thus the users can choose whether to use speech or point on

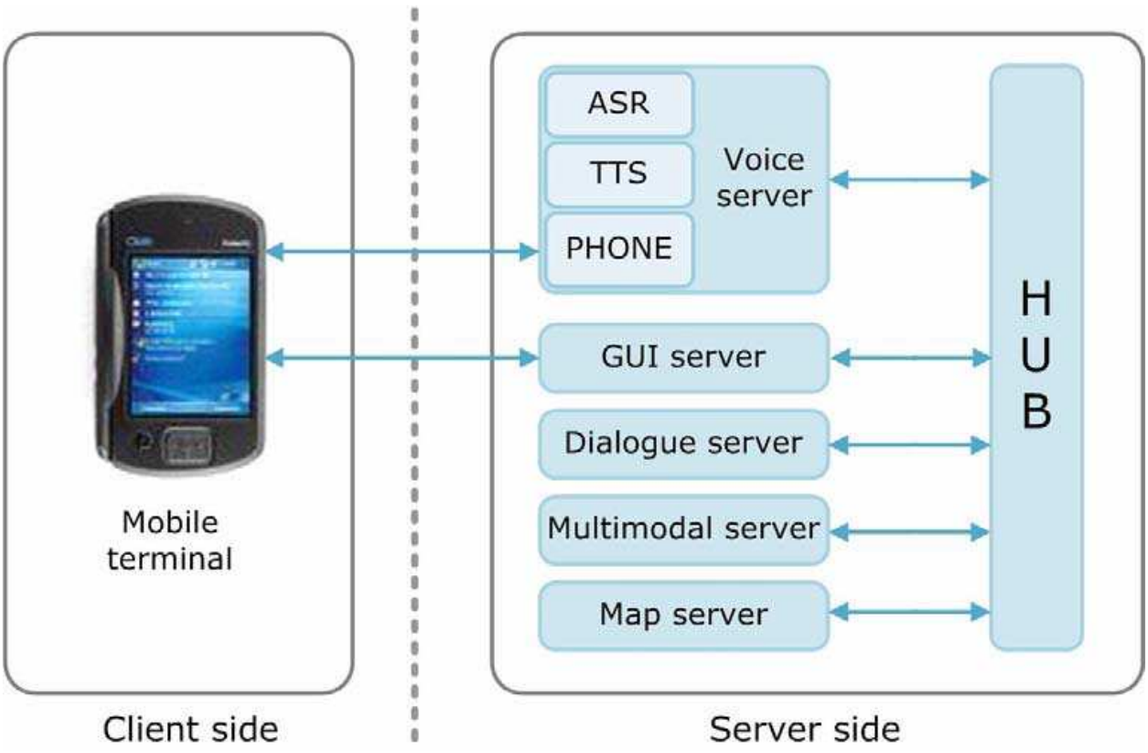

Fig. 4. Multimodal dialogue system software architecture

the map, or even use pointing and talking simultaneously (so-called composite multimodality) to specify the arrival and departure bus stops. The response from the system is presented as both speech and text. More details about the system implementation can be found in (Kvale, et al. 2003b), (Warakagoda, et al. 2003), (Kvale, et al. 2004), (Schie, 2006).
When the client of our multimodal service is started and connected to the server, the main page of the server is presented to the user. This is an overview map of the Oslo area where different sub-areas can be zoomed into, as shown in Figure 5.
Once zoomed, it is possible to get the bus stops in the area displayed. The user has to select a departure bus stop and an arrival bus stop to get the bus route information. The users are not strictly required to follow the steps sequentially. They can for example combine several of them, whenever it makes sense to do so.

Fig. 5. A typical screen sequence for a user with reduced speaking ability. 1) Overview map: The user taps on the submap (the square) for Fornebu. 2) The user says ”next bus here Jernbanetorget” and taps on bus stop Telenor. 3) The system does not recognize the arrival bus stop. Therefore the user selects it by using pen. But first the user taps on the zoom-out button to open the overview map. 4) The user taps on the submap where the bus stop Jernbanetorget lies. 5) The user taps on the bus stop Jernbanetorget. 6) The user can read the bus information.

Both tapping and speech can be used in all operations including navigation and selecting bus stops. Thus the user scenarios can embrace all the possible combinations of pointing and speech input. The received bus route information is presented to the user as text in a textbox and this text is also read aloud by synthetic speech, as illustrated in figure 5.

Our service provides both non-coordinated simultaneous inputs (i.e. the speech and pointing inputs are interpreted one after the other in the order that they are received) and *composite* inputs (i.e. the speech and pointing inputs at the "same time" are treated as a single, integrated compound input by downstream processes), as defined by World Wide Web Consortium (W3C, 2003). Users can also communicate with our service monomodally, i.e. by merely tapping on the touch sensitive screen or by speech only. The multimodal inputs can be combined in several ways, for instance:

- The user utters the name of the arrival bus stop and points at another bus stop on the map, e.g.: "I want to go from Jernbanetorget to here"
- The user points at two places on the screen while saying: "When does the next bus leave from here to here".

In both scenarios above the users point at a bus stop within the same time window as they utter the underlined word, "here". In order to handle such inputs, we defined an asymmetric time window within which speech and tapping are treated as a composite input if:

A. ASR is completed within 3 seconds after a tapping is registered (Δt_{tap} = 3 s)
B. Pointing is registered within 0.85 second after ASR is completed (Δt_{tap} = 0.85 s)

where registration of tapping is instantaneous and the speech recognition is completed at the end point of the speech signal, as illustrated in Figure 6.

In order to handle two taps on the screen within the same utterance, an integration algorithm that uses two such time windows is employed (Warakagoda, et al. 2003).

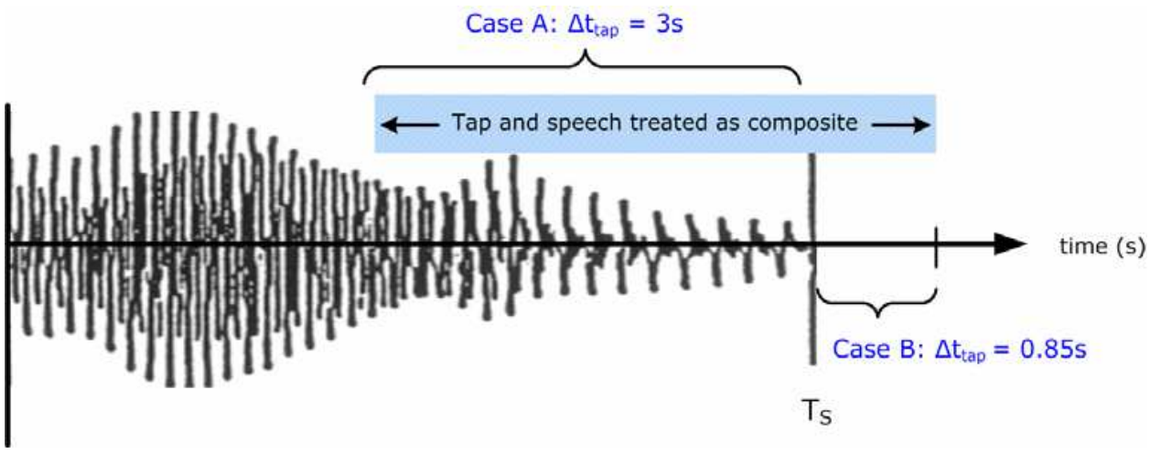

Fig. 6. Example of composite tap and speech inputs. At time T_s the end point of the speech signal is detected and ASR is completed. The blue area illustrates the asymmetric time window around T_s where a tap is interpreted as composite with speech. In case A a tap within a timeframe of maximum 3 seconds before T_s is composite with speech. In case B a tap within 0.85 seconds after T_s is composite with speech.

6. User evaluations

Since the multimodal system gives the users a range of possible input and output alternatives we expect that the service will prove useful for normal-functioning users as well as for many different types of disabled users, such as:

- Persons with impaired hearing or speaking problems who will prefer the pointing interaction.
- Blind persons who will only use the pure speech-based interface
- Users with reduced speaking ability who will use a reduced vocabulary while pointing at the screen.

6.1 Introducing the multimodal for new users

The multimodal interaction pattern was new to the test users and it was necessary to explain this functionality to them. In a user experiment with able-bodied persons we discovered that different introduction formats (video versus text) had a noticeable effect on user behaviour and how new users actually interacted with the multimodal service (Kvale, et al., 2003b). Users who had seen a video demonstration used simultaneous pen and speech input more often than users who had had a text only introduction even if the same information was present in both formats. In our user experiments, 9 out of 14 subjects who had seen the video demo applied simultaneous pen and speech input instantly.

We therefore applied two different strategies in the introduction for the disabled test persons:

- For the scenario-based evaluation we produced an introduction video showing the three different interaction patterns: Pointing only, speaking only, and a combination of pointing and speaking. We did not subtitle the video, so deaf people had to read the information on a text sheet.
- For the in-depth evaluation of the dyslectic user and the aphasic user we applied so-called model based learning, where a trusted supervisor first showed how he used the service and carefully explained the functionality.

Since disabled users often have low self confidence we tried to create a relaxed atmosphere and we spent some time having an informal conversation before the persons tried out the multimodal service. In the scenario-based evaluations only the experiment leader and the test person were present. The in-depth evaluations were performed in cooperation with Bredtvet Resource Centre, a Norwegian national resource centre for special education, representing interdisciplinary expertise within the field of speech, language and communication disorders (Bredtvet, 2009). In the in-depth evaluations the test persons brought relatives with them.

The dyslectic user had his parents with him, while the aphasic user was accompanied by his wife. The evaluation situation may still have been perceived as stressful for them since two evaluators and two speech therapists were watching. This stress factor was especially noticeable in the young dyslectic.

6.2 Scenario-based evaluation

A qualitative scenario-based evaluation followed by a questionnaire was carried out for five disabled users. The goal was to study the acceptance of the multimodal service by the disabled users.

The users were recruited from "Telenor Open Mind", which is a job training programme offering physically disabled people a unique chance for employment (Telenor, 2009). They were in their twenties with an education of 12 years or more. The disabilities of the five users are:

- Muscle weaknesses in hands
- Severe hearing defect and a mild speaking disfluency
- Wheelchair user with muscular atrophy affecting the right hand and the tongue
- Low vision
- Motor control disorder and speech disfluency.

The scenario selected for this evaluation involved finding bus route information for two given bus stops. The users had to complete the task in three different manners: By using pen only, speech only and by using both pen and speech. The tests were carried out in a quiet room with one user at a time. All the test persons were able to complete the tasks in at least one manner:

- They were used to pen-based interaction with PDAs so the pen-only interaction was easy to understand and the test users accomplished the task easily. Persons with muscle weaknesses in hands or with motor control disorder demanded the possibility of pointing at a bigger area around the bus stops. They also suggested that it might be more natural to select objects by drawing small circles than by making a tap, see also (Kvale et al., 2005). The person with hearing defects and speaking disfluency preferred the pen only interaction.
- The speech only interaction did not work properly, partly because of technical problems with the microphone and speech recogniser and partly due to user behaviour such as low volume and unclear articulation.
- The multimodal interaction was the last scenario in the evaluation. Hence some persons had to have this functionality explained to them again before trying to perform this task. The persons with muscular atrophy combined with some minor speaking problems had great benefit from speaking short commands or phrases while pointing at the maps.

In the subsequent interviews all users expressed a very positive attitude to the multimodal system and they recognized the advantages and the potential of such systems (Kristiansen, 2004), (Kvale & Warakagoda, 2005), (Kvale et al. 2005), (Kvale & Warakagoda, 2008).

6.3 In-depth evaluation of a severe dyslectic test user

Dyslexia causes difficulties learning to read, write and spell. Short-term memory, concentration, personal organisation and sequencing may be affected. About 10% of the population may have some form of dyslexia, and about 4% are regarded as severely dyslexic (Dyslexia, 2009).

Our dyslectic test person was fifteen years old and had severe dyslexia. He could, for instance, not read the destination names on the buses. Therefore he was very uncertain and had low self-confidence. He was not familiar with the Oslo area. Thus we spent more than an hour discussing, explaining and playing with the multimodal system. The dyslectic sat beside his trusted supervisor/speech therapist who showed him how to ask by speech only for bus information to travel from "Telenor" to "Jernbanetorget". The speech therapist repeated and rephrased the query: "Bus from "Telenor" to Jernbanetorget" at least five times, and the dyslectic was attentive.

However, when we asked the dyslectic test person to utter the same query, he did not remember what to ask for. Therefore we told him to just say the names of the two bus stops: "From Telenor to Jernbanetorget". He had, however, huge problems remembering and pronouncing these names, especially "Jernbanetorget" because it is a long word. Hence we simplified the task to asking for the bus route information: "From Telenor to Tøyen", which was easier for him. But he still had to practise a couple of times to manage to remember and pronounce the names of these two bus stops.

Then he learned to operate the PDA and service with pointing only. After some training, he had no problem using this modality. He quickly learned to navigate between the maps by pointing at the "zoom"-button. The buttons marked F and T (see figure 5) were intuitively recognised as From station and To station respectively.

Then we told him that it was unnecessary to formulate full sentences when talking to the system, one word or a short phrase was enough to trigger the dialogue system. He then hesitatingly said "Telenor". The system responded with "Is Telenor your from bus stop?", and he answered "yes". In situations where the system did not understand his confirmation input, "yes", he immediately switched to pointing at the "yes" alternative on the screen (he had no problem reading short words). If the bus stop had a long name he would find it on the map and select it by pen instead of trying to use speech.

Finally we introduced the composite multimodal input functionality. We demonstrated queries as: "from here to here" simultaneously tapping the touch screen and saying "here". The dyslectic then said "from here" and pointed at a bus stop shortly afterwards. Then he touched the 'zoom out' button and changed map. In this map he pointed at a bus stop and then said: "to here". This request was correctly interpreted by the system which responded with the bus route information. Both the speech therapists and the parents were really surprised by how well the young severe dyslectic boy managed to use and navigate this system. His father concluded: "When my son learned to use this navigation system so quickly – it must be really simple!"

6.4 In-depth evaluation of an aphasic test user

Aphasia refers to a disorder of language following acquired brain damage, for example, a stroke. Aphasia denotes a communication problem, which means that people with aphasia have difficulty expressing thoughts and understanding spoken words, and they may also have trouble reading, writing, using numbers or making appropriate gestures.

About one million Americans suffer from aphasia (Brody, 1992). There is no official statistics for the number of aphasic persons in Norway. Approximately 12000 people suffer a stroke every year and it is estimated that about one third of these result in aphasia. In addition, accidents, tumours and inflammations may lead to aphasia, giving a total of about 4000-5000 new aphasia patients every year in Norway.

Our test person suffered a stroke five years ago. Subsequently he could only speak a few words and had paresis in his right arm and leg. During the first two years he had the diagnosis global aphasia, which is the most severe form of aphasia. Usually this term applies to persons who can only say a few recognizable words and understand little or no spoken language. Our test person is no longer a typical global aphasic. He has made great progress, and now he speaks with a clear pronunciation and prosody. However, his vocabulary and sentence structure are still restricted, and he often misses the meaningful words – particularly numbers, important verbs and nouns, such as names of places and

persons. He compensates for this problem by a creative use of body language and by writing numbers. He sometimes writes the first letter(s) of the missing word and lets the listener guess what he wants to express. This strategy worked well in our communication. He understands speech well, but has problems interpreting composite instructions. He is much better at reading and comprehending text than at expressing what he has read.
Because of his disfluent speech, characterized by short phrases, simplified syntactic structure, and word finding problems, he can be classified as a Broca's aphasic, although his clear articulation does not completely fit this classification.
He is interested in technology and has used a text-scanner with text-to-speech synthesis for a while. He knew Oslo well and was used to reading maps. He very easily learned to navigate with the pen pointing. He also managed to read the bus information appearing in the text box on the screen, but he thought that the text-to-speech reading of the text helped his comprehension.
His first task in the evaluation was to get bus information for the next bus from “Telenor” to “Tøyen” by speaking to the service. These bus stops are on different maps and the route implies changing buses. Therefore, for a normal user, it is much more efficient to ask the question than pointing through many maps and zooming in and out. But he did not manage to remember and pronounce these words one after the other.
However, when demonstrated, he found the composite multimodal functionality of the service appealing. He started to point at the from-station while saying “this”. Then he continued to point while saying “and this” each time he pointed – not only at the bus stops but also at function buttons such as “zoom in” and when shifting maps. It was obviously natural for him to talk and tap simultaneously. Notice that this interaction pattern may not be classified as a composite multimodal input as defined by W3C, because he provided exactly the same information with speech and pointing. We believe, however, that if we had spent more time in explaining the composite multimodal functionality he would have taken advantage of it.
He also tried to use the public bus information service on the web. He was asked to go from “Telenor” to “Tøyen”. He tried, but did not manage to write the names of the bus stops. He claimed that he might have managed to find the names in a list of alternatives, but he would probably not be able to use this service anyway due to all the problems with reading and writing. The telephone service was not an alternative for him at all because he was not able to pronounce the bus stop names. But he liked the multimodal tap and talk interface very much and spontaneously characterised it as ”Best!”, i.e. the best alternative for him to get the information needed.

7. Discussion

In this chapter we have shown that multimodal human-computer interfaces offer the users the opportunity to choose the most natural interaction pattern for the actual application and context of use. If the preferred mode fails in a certain context or task, users may switch to a more appropriate mode or they can combine modalities. For able-bodied users multimodal interfaces imply enhanced user-friendliness and flexibility in the use of the services, whereas for the disabled users this is a means by which they can compensate for their impaired communication mode.
We have developed a flexible speech centric composite multimodal interface to a map-based information service on handheld mobile terminals such as wireless personal digital assistant

(PDA) devices and 3rd generation mobile phones (3G/UMTS/HSPA). Both tapping and speech can be used in all operations including navigation and selecting bus stations. To the best of our knowledge, our multimodal interface is still the only system with the capability of handling composite inputs consisting of two taps within same spoken utterance.
This user interface proved to be useful for people with different types of disabilities, from muscular atrophy combined with some minor speaking problems, to dyslexia and aphasia.
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.
These qualitative evaluations of how users with reduced ability interacted with the multimodal interface are by no means statistically significant. We are aware that there is a wide variation among aphasics, and even the performance of the same person may vary from one day to the next. Still, it seems reasonable to generalise from our observations and claim that for severe dyslectics and certain groups of aphasics a multimodal interface can be the only useful interface to public information services such as bus timetables. Since most aphasics have severe speaking problems they probably will prefer to use the pointing option, but our experiment indicates that they may also benefit from the composite multimodality since they can point at the screen while uttering simple supplementary words.
Our speech-centric multimodal service allowing all combinations of speech and pointing has therefore the potential of benefiting non-disabled as well as disabled users, and thereby achieving the goal of a common design for all.

8. Conclusion

In this chapter we have demonstrated how multimodal human-computer interfaces are able to combine different input signals, extract the combined meaning from them, find requested information and present the response in the most appropriate format. Multimodal interfaces offer the users an opportunity to choose the most natural interaction pattern depending on the actual task to be accomplished, the context, and their own preferences and abilities. Hence, multimodal user interfaces have the potential to fulfil the requirements and guidelines for Universal Design.

9. Acknowledgements

We would like to express our thanks to Tone Finne, Eli Qvenild and Bjørgulv Høigaard at Bredtvet Resource Centre for helping us with the user evaluation and for valuable discussions and cooperation. We are grateful to our colleagues Ragnhild Halvorsrud, Jon Emil Natvig and Gunhild Luke at Telenor for their inspiration and help.

10. References

Almeida, L. et al. (2002 a). Implementing and evaluating a multimodal and multilingual tourist guide. In: *Proc. International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems*, van Kuppevelt, J. et al. (eds.) 2002., pp. 1–7, Copenhagen, Denmark

Almeida, L. et al, (2002 b). The MUST guide to Paris., Implementation and expert evaluation of a multimodal tourist guide to Paris, In: *Proc. ISCA (International Speech Communication Association) tutorial and research workshop on Multi-modal dialogue in Mobile environments (IDS2002),* Kloster Isree, Germany

Bolt, R. A. (1980). Put That There: Voice and Gesture at the Graphics Interface, *Proceedings of the 7th annual conference on Computer graphics and interactive techniques,* 14(3), pp 262-270. ISBN:0-89791-021-4. Seattle, Washington, United States.

Beskow, J. et al. (2002), Specification and Realisation of Multimodal Output in Dialogue System, *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, pp.181-184. Denver, USA, 2002

Bredtvet (2009). Bredtvet Resource Centre. URL, http://www.statped.no/bredtvet, Accessed: 01.11.2009

Brody, J.E. (1992). When brain damage disrupts speech, In: *The New York Times Health Section*, p. C13, June 10, 1992.

Bühler D. & Minker, W. (2005). The SmartKom Mobile Car Prototype System for Flexible Human-Machine Communication, In: *Spoken Multimodal Human-Computer Dialogue in Mobile Environments*, Springer, Dordrecht (The Netherlands), 2005

Dyslexia Action, URL, http://www.dyslexiaaction.org.uk/, Accessed: 01.11.2009

ETSI (2003). Human Factors (HF); Multimodal interaction, communication and navigation guidelines. Sophia Antipolis, 2003. The European Telecommunications Standards Institute (ETSI EG 202 191).

ETSI (2009). Human Factors (HF); Guidelines for ICT products and services; "Design for all", The European Telecommunications Standards Institute (ETSI) EG 202 191 v1.2.2., (2009-03)

Galaxy (2007). Galaxy communicator. URL, http://communicator.sourceforge.net/, Accessed 24.05.2007.

GSM (2009). GSM Arena, URL, http://www.gsmarena.com/, Accessed: 01.11.2009

Gustafson, J. et al. (2000). Adapt- A Multimodal Conversational Dialogue System In An Apartment Domain, *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP 2000)*, Vol. II, pp.134-137. Beijing, China.

Hui, P.Y. &. Meng, H.M. (2006). Joint Interpretation of Input Speech and Pen Gestures for Multimodal Human Computer Interaction, *Proceedings of. INTERSPEECH – ICSLP'2006*, pp. 1197-1200. Pittsburgh, USA.

Hurtig, T. (2006). A Mobile Multimodal Dialogue System for Public Transportation Navigation Evaluated, *Proceedings of the MobileHCI'06*, September, 12–15, 2006, Helsinki, Finland.

Johnston, M.; Srinivas, B. & Gunaranjan, V. (2001). MATCH: multimodal access to city help. *Proceedings of the Automatic Speech Recognition and Understanding Workshop*, Madonna Di Campiglio, Trento, Italy

Johnston, M. et al. (2002). Multimodal language processing for mobile information access, *Proceedings of the ICSLP-2002*, pp. 2237-2240. 2002.

Karpov, A.; Ronzhin, A. & Cadiou, A. (2006). A Multi-Modal System ICANDO: Intellectual Computer AssistaNt for Disabled Operators, *Proceedings of INTERSPEECH - ICSLP 2006*, pp. 1998-2001, Pittsburgh, USA

Kristiansen, M. (2004). Evaluering og tilpasning av et multimodalt system på en mobil enhet, Master thesis NTNU (in Norwegian), 2004.

Kumar, S.; Cohen, P.R. & Levesque, H.J. (2000). The adaptive agent architecture: achieving fault-tolerance using persistent broker teams, *Proceedings of the Fourth International Conference on MultiAgent Systems*, pp. 159-166. 2000.

Kvale, K; Warakagoda, N.D. & Knudsen, J.E. (2003a), Speech centric multimodal interfaces for mobile communication systems, In: *Telektronikk.* Vol. 99. No. 2, pp. 104-117. ISSN 0085-7130

Kvale, K.; Rugelbak J. & Amdal, I. (2003b). How do non-expert users exploit simultaneous inputs in multimodal interaction?, *Proceedings of International Symposium on Human Factors in Telecommunication*, pp.169-176, Berlin, Germany.

Kvale, K.; Knudsen, J.E. & Rugelbak, J. (2004). A Multimodal Corpus Collection System for Mobile Applications, *Proceedings of Multimodal Corpora - Models of Human Behaviour for the Specification and Evaluation of Multimodal Input and Output Interfaces*, pp. 9-12, Lisbon, Portugal.

Kvale, K.; Warakagoda N.D. & Kristiansen, M. (2005). Evaluation of a mobile multimodal service for disabled users, *Proceedings of the 2nd Nordic conference on multimodal communication*, pp. 242-255. Gothenburg, Sweden

Kvale, K. & Warakagoda N.D. (2005). A Speech Centric Mobile Multimodal Service useful for Dyslectics and Aphasics, *Proceedings of the INTERSPEECH - EUROSPEECH'2005*, pp. 461-464. Lisbon, Portugal

Kvale, K. & Warakagoda, N.D., (2008). Speech centric multimodal interfaces for disabled users, In: Technology and Disability, Special Issue: Electronic speech processing for persons with disabilities. AAATE (Association for the advancement of Assistive Technology in Europe). IOS Press Amsterdam, Washington, DC, Tokyo, Volume 20, No. 2, 2008.pp. 87-95, ISSN 1055-4181

NC, (1997). The Principles of Universal Design, Version 2.0. Raleigh. North Carolina State University. URL, http://www.design.ncsu.edu/cud/about_ud/docs/use_guidelines.pdf, Accessed: 30/10/09

OOA (2009). OAA, Open Agent Architecture, URL, http://www.ai.sri.com/~oaa, Accessed: 30.10.2009

Oviatt, S. et al. (2000). Designing the user interface for multimodal speech and gesture applications: State-of-the-art systems and research direction, In: Human Computer Interaction, vol. 15, no. 4, pp. 263-322. 2000.

Oviatt, S. (2000). Multimodal system processing in mobile environment, In: Proc. of the Thirteenth Annual ACM Symposium on User Interface Software Technology (UIST'2000), ACM: New York, N.Y., 21-30. 2000.

Oviatt, S.; Coulston R. & Lunsford, R. (2004). When Do We Interact Multimodally? Cognitive Load and Multimodal Communication Patterns. In: Proc. of ICMI, 2004.

Smartkom, (2007). SMARTCOM – Dialog-based Human-Technology Interaction by Coordinated Analysis and Generation of Multiple Modalities, URL, http://www.smartkom.org/start_en.html, Accessed: 27.10.2009

Schie, T. (2006). Mobile Multimodal Service for a 3G-terminal, M.S. thesis, Norwegian University of Science and Technology, 2006.

Trafikanten (2009). "Trafikanten Reiseplanleggeren", URL, http://www.trafikanten.no, Accessed: 27.10.2009

Telenor (2009). Telenor Open Mind, URL, http://www.telenor.com/en/people-and-opportunities/programme-for-the-physically-challenged/, Accessed: 27.10.2009

Wahlster, W. et al. (2001). "SmartKom: Multimodal Communication with a Life-Like Character", *Proceedings of the EUROSPEECH-2001*, pp 1547–1550. Aalborg, Denmark

Warakagoda, N. D.; Lium, A.S. & Knudsen, J.E. (2003). Implementation of simultaneous co-ordinated multimodality for mobile terminals, In: The 1st Nordic Symposium on Multimodal Communication, Copenhagen, Denmark, 2003.

Warakagoda N. D., Lopez J. C. L. and Kvale K, (2008). VOICE TICKETING - Method and system for performing an e-commerce transaction., PCT application publication WO/2008/103054, URL, http://www.wipo.int/pctdb/en/wo.jsp?wo=2008103054, Accessed: 30/10/09

W3C, (2003). Multimodal Interaction Requirements, NOTE 8 January 2003, URL, http://www.w3.org/TR/2003/NOTE-mmi-reqs-20030108/, Accessed: 27.10.2009.

W3C, (2009). EMMA: Extensible MultiModal Annotation markup language, W3C Recommendation 10 February 2009. URL, http://www.w3.org/TR/emma/, Accessed: 27.10.2009

WCAG, (2008). Web Content Accessibility Guidelines (WCAG) 2.0. W3C Recommendation 11 December 2008. URL, http://www.w3.org/TR/WCAG20/, Accessed: 02.10.2009

Wang, Ye-Yi. (2001). Robust language understanding in MiPad, *Proceedings of the EUROSPEECH-2001*, pp 1555–1558, Aalborg, Denmark, 2001.

Watanabe, Y. et al. (2007). Semi-synchronous speech and pen input, *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. 2007, pp. IV. 409-412.

Wu, L.; Oviatt, S.L. & Cohen, P. R. (1999). Multimodal Integration – A Statistical View. *IEEE Trans. on Multimedia*, 1 (4), 1999. pp. 334–341. ISSN: 1520-9210

User Interfaces

Edited by Rita Matrai

ISBN 978-953-307-084-1
Hard cover, 270 pages
Publisher InTech
Published online 01, May, 2010
Published in print edition May, 2010

Designing user interfaces nowadays is indispensably important. A well-designed user interface promotes users to complete their everyday tasks in a great extent, particularly users with special needs. Numerous guidelines have already been developed for designing user interfaces but because of the technical development, new challenges appear continuously, various ways of information seeking, publication and transmit evolve. Computers and mobile devices have roles in all walks of life such as in a simple search of the web, or using professional applications or in distance communication between hearing impaired people. It is important that users can apply the interface easily and the technical parts do not distract their attention from their work. Proper design of user interface can prevent users from several inconveniences, for which this book is a great help.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen