Multimodal interaction on mobiles and applications*

Gaspoz Arnaud
University of Fribourg
1700 Fribourg
Switzerland
arnaud.gaspoz@unifr.ch

ABSTRACT

The current generation of mobile phones allows traditional interactions via menus and buttons, but offers a variety of new ways of interacting with the device via the built-in sensors. Multimodality through mobile phones are growing in importance and aims to create applications which are socially acceptable and enjoyable. We present multiple multimodal applications made for pedestrians and impaired people. We also talk about more common applications such as games, photos retrieval system and home entertainment systems. For each of them, we describe which modalities are used and how they are combined. The results are analysed in order to determine if the use of multimodality has benefits. Finally, we conclude that multimodality is very efficient in almost all cases and has a great potential in a near future.

Keywords

Multimodality, mobile application, mobile interaction

1. INTRODUCTION

Mobile phones have passed through several stage during their history. Phone adoption has grown rapidly across the world and nowadays, billion of people have these devices and became reachable anytime, anywhere. Mobile technology has been widely adopted from teenagers sending messages to business people trying to improve their livelihood [1].

The current generation of mobile phones allows traditional interactions via menus and buttons, but offers a variety of new ways of interacting with the device via the built-in sensors such as touch sensitive screens, accelerometers, bluetooth, microphone, camera, etc. With the imminent increase in network bandwidth available to mobile and the rise in the number of applications and services provided by smartphones, we need new techniques for efficiently communicating with the device.

Phone usage depends significantly on both location and social context. This two properties influence the way applications are designed and evaluated. Mobile phones generally have small interfaces with limited input facilities (ie. keyboard and touch) which limit their usage while user is on the go. This problem can be resolved by using multimodality. People in motion can't devote all or any part of their visual resource and it's here where multimodal mobile interactions come into play.

Multimodal interactions through mobile phones are growing in importance. Much difficulty for such a system is to create robust applications which are socially acceptable and enjoyable. The next section of the paper focuses on the types of multimodal applications available on mobile. Some categories are explained in more details by noticing which modalities are used and for what actions [2].

2. TYPES OF APPLICATION

Mobile phones are used in an increasing range of places and contexts such as home, work, street, car and other places. Since many years, several studies have tried to understand mobile phone usage and can be categorized into those that use surveys or self-reports and recent studies that exploit recording capability of mobile phone in order to capture usage behaviour. As stated in [1] which is a large-scale study gathering data continuously from 77 participant's mobiles, the most popular applications can be classified into specific types. The retained applications are the sms, voice call, web, multimedia, clock, camera, email, calendar, voice chat, maps, sport tracker and visual radio. We can noticed that the sms and voice call are clearly the most used applications which is obvious because they are the two primary functions of a mobile phone. The leader in the mobile phone's share market, Android, and its direct concurrent, iOS, both offer the functionality to send messages or to record a new event in the calendar with the voice.

For the remainder of the paper, only multimodal applications are kept. After some research in the domain, a lot of studies are made on developing true multimodal applications. That goes from the everyday usage of mobile phone to search something on the phone or on the Internet to specialized applications made for impaired-people. All re-

*This paper is a deliverable of the Msc research seminar on "Multimodal Interaction on Mobile Devices", organized in 2012 in the department of Informatics of the University of Fribourg: https://diuf.unifr.ch/main/diva/teaching/seminars/seminar-multimodal-interaction-mobiles-devices

lated works found are divided into categories and each part is explained in more details with an example. All concerned applications fit into the following categories: pedestrians, impaired-people, home entertainment, multimedia and game. For each categories, the modalities used are identified and their role in the process are explained.

2.1 Pedestrians

Pedestrians are the first category where applications use multimodality to improve the user experience. Almost all of the mobile user interfaces are designed for a person who stays in place and can concentrate on the screen. When using this device outside, attention and motor ability are limited and users must adapt their use of the mobile phone which lead to a reduced performance. By walking down the street, a person typing a text message or reading text on screen must maintain awareness and avoids obstacles. In the following paragraphs, two evaluations about applications using different modalities designed for pedestrians are presented.

The first experiment evaluates the use of different combinations of modalities, for example only speech as input or using speech with pointing gestures in order to complete some given tasks and are described in [3]. For evaluating the multimodality paradigm in mobile system, particularly in place and contexts that affect interactions, the system allows visitors of a city to interact with a map, for example pan and zoom, to get information on sights and planning a tour from a start to an end. These tasks represent some generic functionalities in today's devices. The evaluation has used 12 subjects and quantitative and qualitative data have been collected through log files, questionnaires, observation and video recording. The experiment includes situational aspects such as sitting in a cafe, sitting in the bus, standing at a viewpoint and walking in a pedestrian zone. The usability has been rated using 5 parameters: convenience, speed, intuitive usability, efficiency and overall acceptance.

The results show that the use of multimodality is accepted by the users and in almost all the cases, performs better than standard interactions. Distractions such as passing pedestrians or reflecting displays due to sunlight underline the benefits of using natural language in order to communicate with the device. Another remark is that task with more complex cognitive load are more quickly performed using voice. Finally, regarding the acceptance of multimodality in situational context, users might have concerns about interacting with some modality when strangers are around and the acceptance decreases with social unfamiliarity of accompanying persons.

The second experiment introduces the term Walking User Interfaces (WUIs) which are GUIs that adapt their layout when user is moving[4]. As previously stated, using a mobile phone during walking impact on the user's ability to read or write text. Based on this fact, mobile phone must be able to infer some user's context by sensing the environment, for example with the accelerometer and to adapt the application by changing the text size or by providing alternative interface. For the evaluation, 30 persons are involved in the experiment and they are asked to interact with the touch screen of the device using their fingers, but no styli or hard button are used. The test assesses the effect of different button sizes or dynamically changing the size of the user interface in order to see if performance are improved when walking. The application is a portable music player on which the users have to scroll through the list to find the desired song and to click on it to launch it.

The results reveal that the time needed in order to accomplish a given task is longer when the interface is very small or very large. This is obvious by the fact that small screen takes more time to pick the song and that the larger text size for the song name requires more scrolling. Another aspect noticed is the use of simpler interface containing only important functions in order to provide a smaller set of actions to the user but with better performance. Nevertheless, WUIs were not best than static interface, but performed better for 10 of the 30 users. The adaptive interface was no slower or more error-prone than standard interfaces. To conclude, the study confirms that adaptively changing the interface size shows only small performance differences.

With this two cases, we see the importance of the evaluation of prototype in order to quantify the performance brought. The voice seems to be a good alternative when user is walking, but there are some issues such as the privacy.

2.2 Impaired people

In this section, we focused on applications made for impaired people. The mobile phone market is moving towards touch screen featured devices which support an inherently visual experience. The interface of new mobile phones doesn't have accessibility support for the visually impaired people. As a result, services such as screen readers or voice commands are developed on the top of the existing applications.

A solution to address the difficulties of using mobile phones for visually impaired people is to introduce multimodality on current devices. This way, the user has the choice of the modality to use for interacting with the system. This freedom is particularly enjoyable for disabled persons. Many researches have been done to improve the accessibility of mobile phones, especially with text-to-speech system, but here we present two others studies which aim to find solutions in order to make life more pleasant for partially sighted people.

The first study introduces an application named Timbremap which is a novel interface for map and floor-plan exploring using audio feedback in order to correct the user's finger on the touch interface. Currently, disabled users rely on text-to-speech systems in order to use map applications. But such systems are limited in transmitting complex informations and spacial positioning. Timbremap can be installed on today's mobile phones with multi-touch capabilities. The audio feedback mechanism allows visually impaired people to navigate through complex indoor layout by creating a mental map of the place. The application targets primarily real-time navigation of indoor environments. The prototype implements a line and an area hinting mode which basically works like explained below: "First, the user hears a repeating audio pattern representing the path segment. As the user's touch shifts to the left, a repeating beep in the right ear is played to indicate that the nearest segment is to the right of the finger. This interface uses mid-tone beeps on

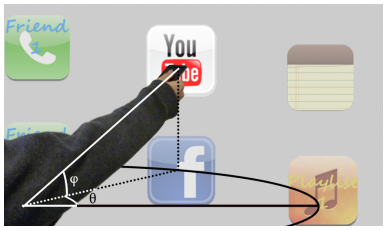

Figure 1: Virtual Shelves uses the theta (θ) and phi (ϕ) angles to index into a shelf of shortcuts.

stereo channels to indicate whether path is to the left or right of the touch point. For up and down adjustments, this uses high-pitched up and low-pitched down spearcons, respectively"[5]. In addition to previously cited audio feedbacks, the application provides sound for points of interest. The user can hold one finger on the point of interest marker and double-click anywhere else with another finger. Then the system gives informations about the marker using the text-to-speech service.

The evaluation involves 6 visually impaired people. The results show that there are no significant difference between the two proposed modes. On average, participants identify shape with 81% accuracy in only 41 seconds per shape. Timbremap conveys complex informations about indoor floorplans allowing users to navigate in that environment.

The second case included here is an interaction technique based on Virtual Shelves aiming to facilitate the access of application shortcuts. As previously mentioned, using a mobile phone for visually impaired people is a difficult task, even more for accessing advanced functions. "Virtual Shelves utilizes an orientation-aware mobile device to determine the theta and phi angles of the user's arm with respect to her body (see Figure 1) and uses these angles to index into a list of shortcuts"[6]. This application can be use on new mobile phones which have an accelerometer and a gyroscope in order to determine its position and orientation.

The experiment involves 13 visually impaired people which have to customize 15 regions with 7 personal shortcuts. The results show that selection error is greater on top regions than the middle and bottom regions. Overall, shortcuts have been accessed with 88.3% accuracy in personal layout settings in only 1.74 seconds. This results are very promising and open some doors for helping disabled people to better integrate our society.

2.3 Home entertainment

In this section, we focus on an exciting domain for multimodal application: home entertainment. Since a few years, mobile phone became much more than a simple device for calling somebody, they are more powerful, embedding sensors which bring new potential such as using mobile phone for controlling home environment. The mobile device can replace multiple remote controls and consequently provides new modalities for inputs and outputs.

In this experiment[7], the prototype uses a high-definition television with a program guide and a Nokia N95 phone which allows remote control. The media center can be manipulated by speaking, for example by saying sentences such as "Show me all children programs". The user have to press a button before telling the command. Then the record is sent to a server which perform the speech recognition with the help of grammar-based speech interface. In addition to speech, gesture is supported via the accelerometer and the patterns are processed by a gesture recognizer which communicates the command to the system. For moving in the menu, the user can simply press the phone keyboard. Finally, the vibration component of the device is used as haptic feedback in order to notify the user that a gesture has been recognized or the recording of a speech has ended.

The study involves 26 participants which have to fill an "expectations" pre-test and a "perception" post-test in order to determine the gap between them. The evaluation covered several points such as speed, naturalness, clearness, error free, usefulness, etc for each modality. The results show that the speech modality is widely accepted as input modality. With an accuracy of 93%, the speech recognition is very effective and we can assume that the speech allows issuing more complex commands than the two other modalities. Concerning the gesture input and the haptic feedback, both of them were expected as good modalities for interaction by many users but were not really enjoyed. We can conclude that gesture can submit more complex tasks than navigation and that haptic feedback should be simple as possible. With this example, we saw that multimodality could be use for an everyday usage and make life easier.

2.4 Multimedia

With the progress in the domain of mobile phones, these devices are now able to process multimedia. With the camera, mobile phones can easily take photos or record videos and watch them again by selecting one in the gallery. The following application allows the user to annotate, index and search pictures on a mobile phone.

The application presented here is named MAMI (Multimodal Automatic Mobile Indexing) and allows annotate pictures stored on a mobile phone and to search and retrieve them via multimodal inputs[8]. Nowadays, it's common to see people taking pictures with their phones and therefore it's become harder to search for a specific photo. Browsing the gallery in order to find a picture can be quite long. MAMI uses multimodality to annotate images and then use inputs from different modalities as queries in order to increase the accuracy of finding the desired picture.

The prototype works as following: At the time of capturing or at a later moment, the user can record for a given picture an audio tag which is indexed in a local database. The audio recording is performed by a push-to-talk method. Afterwards, the system generates descriptors for the audio record and the image itself. The audio processing doesn't use speech recognition, but creates fingerprint of the record which means that this mechanism is speech independent. For the image processing, MAMI uses the Edge Histogram Descriptor which is based on edge-derived features characterizing the picture content. When looking for a picture, the user can submit a speech record or a sample image. In any cases, the system computes corresponding descriptor which

will be compared to all existing descriptors in the database. Finally, the 4 best pictures are returned to the user and the latter will be happy if the expected photo is in the returned set.

The evaluation assesses the retrieval accuracy of using mono modal query (speech or image sample) compared to multimodality (audio and visual feature). The test involves 6 participants and 546 pictures, 91 on average. In order to use multiple modalities to search an image, the system performs fusion techniques at a decision level, which mean that the data are processed independently for each modality and the fusion algorithm is responsible to take the decision. The results show that the speech input outperforms the image analysis for individual modality and the accuracy reaches 94.48% for the audio and 83.87% for the image. Concerning the fusion, the evaluation has tested three multimodal fusion techniques and all have returned better results with on average 95.63%. The drawback of this approach is that she requires an image and an audio tag for each request.

2.5 Game

With dual-core processor and huge amount of memory, even games can be run on current devices. In this last section, we present a game in which the user has to reproduce ordered actions with different modalities.

We will focus now on a game named SensoDroid[9]. The game is based on the well-know Senso or Simon traditional game. The idea of developing a multimodal game comes from the fact that the market for electronic entertainment on mobile phone is constantly growing. Recent years have shown the success of new interaction techniques such as the Wiimote, Microsoft Kinect and Sony Move. This project follows this trend by using the embedded sensors of mobile phones such as accelerometer, absolute position transducer and touch screen. The concept of the game is to integrate multimodality as interaction techniques. The goal is to memorize and to reproduce within a limited time a serie of sensor actions in the correct order. Through the menu, the user can launch the game which is composed of two phases: the Memorize Interactions and the Replay Interactions phase. The first one generates a random sequence of interactions represented by a text and graphics. During the second phase, the user reproduces the previously shown interactions in the correct order.

The prototype uses the Java programming language in order to develop a game for the Android platform. The system is designed to interpret the raw data into user inputs and generates the corresponding actions. The accelerometer is used to detect pitch and rolling actions. The compass actions are captured by the built-in digital compass and are needed to test if the mobile phone is aligned with respect to the cardinal directions. Microphone processes the auditory inputs and the touch screen is responsible for simple events such as touch and double touch and also to register wiping gestures. As feedback, a simple vibration and acoustic signal are provided in order to signal the user that an action has been recognized. The evaluation results show that a game using multimodal interaction techniques could be very attractive.

By the hardware aspect of current mobile phones, multimodality is well suited for interacting with the devices, especially for recreational aspects.

3. CONCLUSION

In this paper, we presented multiple multimodal applications available on mobile phones. These applications have been categorized and described in details in order to have a taste of how they work. The emphasis was put on which modalities are used and how they are combined to improve the user experience and/or the performance for achieving a given task. We started by talking about specialized applications for pedestrians and for impaired people to more general purpose applications such as software for home entertainment, photo retrieval and games.

The results of the different experiments showed all their importance in order to determine if the use of a modality performed better than the standard input. Performance is not always the only criteria for measuring multimodal applications, we can take as example the visually impaired people which can use services that can not be the case with touch screen input and output. Almost all studies reveal a great potential of using multimodality in future mobile phone applications.

4. REFERENCES

[1] Daniel Gatica-Perez Trinh Minh Tri Do, Jan Blom. Smartphone usage in the wild: a large-scale analysis of applications and context. *In Proc. ICMI*, pages 353–360, November 2011.

[2] Stephen Brewster Julie R. Williamson, Andrew Crossan. Multimodal mobile interactions: usability studies in real world settings. *In Proc. ICMI*, pages 361–368, November 2011.

[3] Matthias Merdes Rainer Malaka Matthias Jöst, Jochen Häußler. Multimodal interaction for pedestrians: an evaluation study. *In Proc. IUI*, pages 59–66, January 2005.

[4] Ian E. Smith Shaun K. Kane, Jacob O. Wobbrock. Getting off the treadmill: evaluating walking user interfaces for mobile devices in public spaces. *In Proc. MobileHCI*, pages 109–118, September 2008.

[5] Ashvin Goel Eyal de Lara Khai N. Truong Jing Su, Alyssa Rosenzweig. Timbremap: enabling the visually-impaired to use maps on touch-enabled devices. *In Proc. MobileHCI*, pages 17–26, September 2010.

[6] Khai N. Truong Frank Chun Yat Li, David Dearman. Leveraging proprioception to make mobile phones more accessible to users with visual impairments. *In Proc. ASSETS*, pages 187–194, October 2010.

[7] Juho Hella Tomi Heimonen Jaakko Hakulinen Erno Mäkinen Tuuli Laivo Hannu Soronen Markku Turunen, Aleksi Melto. User expectations and user experience with different modalities in a mobile phone controlled home entertainment system. *In Proc. MobileHCI*, (31), September 2009.

[8] Nuria Oliver Xavier Anguera, JieJun Xu. Multimodal photo annotation and retrieval on a mobile phone. *In Proc. MIR*, pages 188–194, October 2008.

[9] Christian Geiger Thomas Reufer, Martin Panknin. Sensodroid - multimodal interaction controlled mobile gaming. *In Proc. HC*, pages 32–36, December 2010.