Mobility and Multimodal User Interfaces

Christopher J. Pavlovski
IBM Corporation, Australia

Stella Mitchell
IBM T. J. Watson Research, USA

INTRODUCTION

Traditional user interface design generally deals with the problem of enhancing the usability of a particular mode of user interaction, and a large body of literature exists concerning the design and implementation of graphical user interfaces. When considering the additional constraints that smaller mobile devices introduce, such as mobile phones and PDAs, an intuitive and heuristic user interface design is more difficult to achieve.

Multimodal user interfaces employ several modes of interaction; this may include text, speech, visual gesture recognition, and haptics. To date, systems that employ speech and text for application interaction appear to be the mainstream multimodal solutions. There is some work on the design of multimodal user interfaces for general mobility accommodating laptops or desktop computers (Sinha & Landay, 2002). However, advances in multimodal technology to accommodate the needs of smaller mobile devices, such as mobile phones and portable digital assistants, are still emerging.

Mobile phones are now commonly equipped with the mechanics for visual browsing of Internet applications, although their small screens and cumbersome text input methods pose usability challenges. The use of a voice interface together with a graphical interface is a natural solution to several challenges that mobile devices present. Such interfaces enable the user to exploit the strengths of each mode in order to make it easier to enter and access data on small devices. Furthermore, the flexibility offered by multiple modes for one application allows users to adapt their interactions based on preference and on environmental setting. For instance, hands-free speech operation may be conducted while driving, whereas graphical interactions can be adopted in noisy surroundings or when private data entry, such as a password, is required in a public environment.

In this article we discuss multimodal technologies that address the technical and usability constraints of the mobile phone or PDA. These environments pose several additional challenges over general mobility solutions. This includes computational strength of the device, bandwidth constraints, and screen size restrictions. We outline the requirements of mobile multimodal solutions involving cellular phones. Drawing upon several trial deployments, we summarize the key designs points from both a technology and usability standpoint, and identify the outstanding problems in these designs. We also outline several future trends in how this technology is being deployed in various application scenarios, ranging from simple voice-activated search engines through to comprehensive mobile office applications.

BACKGROUND

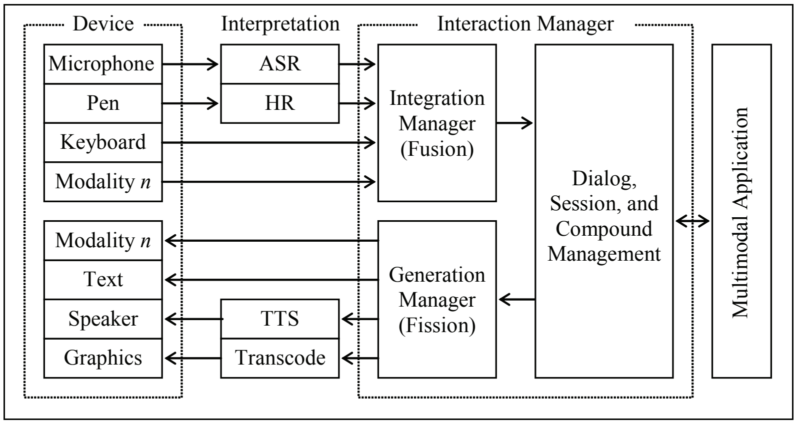
Multimodal interaction is defined as the ability to interact with an application using multiple sensory channels (i.e., tactile, auditory, visual, etc.). For example, a user could provide input by speaking, typing on a keypad, or handwriting, and receive the subsequent response in the form of an audio prompt and/or a visual display. Useful multimodal applications can cover a broad spectrum including tightly synchronized, loosely synchronized, and complementary modes of operation. Synchronization behavior must be defined both for input (the way in which input from separate modes is combined) and for output (the way in which input from one mode is reflected in the output modes). The W3C distinguishes several types of multimodal synchronization for input as follows (W3C, 2003a):

- **Sequential:** Two or more input modalities are available, but only a single modality is available at any given time.
- **Simultaneous:** Allows input from more than one modality at the same time, but each input is acted upon separately in isolation from the others.
- **Composite:** Provides for the integration of input from different modes into one single request.

A general framework for multimodal systems is depicted in Figure 1. This diagram elaborates further on several fundamentals positioned by W3C.

The interaction manager is responsible for combining multiple requests, dialog management, and synchronization. The function of receiving and combining multiple inbound

Figure 1. Multimodal framework



requests is the responsibility of the *integration manager* subcomponent. Conversely, the generation manager is responsible for distributing multimodal output to all of the respective output channels (modes) via an interpretation layer, which may involve text to speech (TTS) conversion or transcoding of graphical content to accommodate the needs of the target modality. Earlier work in multimodal systems referred to the integration tasks relating to composition and decomposition of requests as *fusion* and *fission* respectively (Coutaz, Nigay, & Salber, 1993).

Speech-based telephone interfaces currently available in the commercial market commonly use varying levels of directed dialog. Directed dialog, as the name implies, employs a style of system prompts that helps to "direct" the user in what to say next. Users are often presented with spoken menu options from which they can make a selection, thus navigating in a controlled manner until the task is completed. Much of the naturalness and power of speech is undermined when the application relies too heavily on the use of directed dialogs. A Natural Language speech interface, which allows the user to phrase their request in a wide variety of ways, reduces the cognitive load since there are no commands to memorize or hierarchies to navigate. A mixed-initiative interface allows the user to share control over the direction of the dialog, making the interaction more efficient for the user.

Device manufacturers can install specialized software or firmware on handsets to enable distributed speech recognition (DSR). DSR technology digitizes the speech signal and sends it over an error-protected data channel to a speech recognizer on a server. Thus the processing is distributed between a terminal client and a server. The accuracy of speech recognition can be better when using DSR because the degradations associated with sending speech over the mobile network, such as low bit rate speech coding and channel transmission errors, are avoided. In addition, DSR allows for the implementation of a multimodal, speech and data, application on devices which do not support simultaneous voice and data connections.

MULTIMODAL TECHNOLOGY IN MOBILITY SYSTEMS

Multimodal applications for small mobile devices must overcome several technical challenges; these include device capability and computational strength, functional and bandwidth constraints of the network, and limitations of the user interface. Present work in multimodality is focused upon the support of two modes of interaction, most typically data (graphics or text) and speech (Kondratova, 2004; Pavlovski, Lai, & Mitchell, 2004a; Hastie, Johnston, & Ehlen, 2002; Kvale, Warakagoda, & Knudsen, 2003). Future trends support additional modes that include visual gesture, lip reading, and haptic responses. In this section we present advances in multimodality supporting speech and data entry, outlining the current state of the technology used and the outstanding challenges yet to be addressed.

Multimodal Architectures

Due to the need to support two or more modes of input and output, the solutions to support multimodal systems are more complex than unimodal systems. Additional capabilities are required to support composite input and output requests, manage multiple application states, and perform session management between devices and the multimodal application services.

There are fundamentally two architectural approaches to constructing multimodal solutions for mobile devices such as mobile phones and PDAs. The most widely investigated architecture appears to involve deployment of an application

Related Content