Chapter 2.36
Speech-Centric Multimodal User Interface Design in Mobile Technology

Dong Yu
Microsoft Research, USA

Li Deng
Microsoft Research, USA

ABSTRACT

Multimodal user interface (MUI) allows users to interact with a computer system through multiple human-computer communication channels or modalities. Users have the freedom to choose one or more modalities at the same time. MUI is especially important in mobile devices due to the limited display and keyboard size. In this chapter, we provide a survey of the MUI design in mobile technology with a speech-centric view based on our research and experience in this area (e.g., MapPointS and MiPad). In the context of several carefully chosen case studies, we discuss the main issues related to the speech-centric MUI in mobile devices, current solutions, and future directions.

INTRODUCTION

In recent years, we have seen steady growth in the adoption of mobile devices in people's daily lives as these devices become smaller, cheaper, more powerful, and more energy-efficient. However, mobile devices inevitably have a small display area, a tiny keyboard, a stylus, a low speed (usually less than 400 million instructions per second) central processing unit (CPU), and a small amount (usually less than 64MB) of dynamic random-access memory. Added to these limitations is the fact that mobile devices are often used in many different environments, such as dark and/or noisy surroundings, private offices, and meeting rooms. On these devices, the traditional *graphical user interface* (GUI)-centric design becomes far less effective than desired. More efficient and easy-

to-use user interfaces are in urgent need. The *multimodal user interface* (MUI), which allows users to interact with a computer system through multiple channels such as speech, pen, display, and keyboard, is a promising user interface in mobile devices.

Multimodal interaction is widely observed in human-human communications where senses such as sight, sound, touch, smell, and taste are used. The research on multimodal human-computer interaction, however, became active only after Bolt (1980) proposed his original concept of "Put That There." Since then, a great amount of research has been carried out in this area (Bregler, Manke, Hild, & Waibel 1993; Codella, Jalili, Koved, Lewis, Ling, Lipscomb, et al., 1992; Cohen, Dalrymple, Moran, Pereira, Sullivan, Gargan, et al., 1989; Cohen, Johnston, McGee, Oviatt, Pittman, Smith, et al., 1997; Deng & Yu, 2005; Fukumoto, Suenga, & Mase, 1994; Hsu, Mahajan, & Acero 2005; Huang, Acero, Chelba, Deng, Droppo, Duchene, et al., 2001; Neal & Shapiro, 1991; Pavlovic, Berry, & Huang, 1997; Pavlovic & Huang, 1998; Vo, Houghton, Yang, Bub, Meier, Waibel, et al., 1995; Vo & Wood, 1996; Wang, 1995). Importantly, the body of this research work pointed out that MUIs can support flexible, efficient, and powerful human-computer interaction.

With an MUI, users can communicate with a system through many different input devices such as keyboard, stylus, and microphone, and output devices such as graphical display and speakers. MUI is superior to any single modality where users can communicate with a system through only one channel. Note that using an MUI does not mean users need to communicate with the system always through multiple communication channels simultaneously. Instead, it means that users have freedom to choose one or several modalities when communicating with the system, and they can switch modalities at any time without interrupting the interaction. These characteristics make the MUI easier to learn and use, and is preferred by users in many applications that we will describe later in this chapter.

MUI is especially effective and important in mobile devices for several reasons. First, each modality has its strengths and weaknesses. For this reason, single modality does not permit the user to interact with the system effectively across all tasks and environments. For example, speech UI provides a hands-free, eyes-free, and efficient way for users to input descriptive information or to issue commands. This is very valuable when in motion or in natural field settings. Nevertheless, the performance of speech UI decreases dramatically under noisy conditions. In addition, speech UI is not suitable when privacy and social condition (e.g., in a meeting) is a concern. Pen input, on the other hand, allows users to interact with the system silently, and is acceptable in public settings and under extreme noise (Gong, 1995; Holzman, 1999). Pen input is also the preferred way for entering digits, gestures, abbreviations, symbols, signatures, and graphic content (Oviatt & Olsen, 1994; Suhm, 1998). However, it is impossible for the user to use pen input if he/she is handicapped or under "temporary disability" (e.g., when driving). MUI, on the other hand, allows users to shift between modalities as environmental conditions change (Holzman, 1999), and hence, can cover a wider range of changing environments than single-modal user interfaces.

Second, different modalities can compensate for each other's limitations and thus provide users with more desirable experience (Deng & Yu, 2005; Oviatt, Bernard, & Levow, 1999; Oviatt & vanGent, 1996; Suhm, 1998). For example, the accuracy of a resource-constrained, midsized vocabulary speech recognizer is low given the current speech technology. However, if the speech recognizer is used together with a predictive T9 (text on 9 keys) keyboard, users can greatly increase the text input throughput compared with using the speech modality or T9 keyboard alone (Hsu et al., 2005). The gain is obtained from the mutual disambiguation effect, where each error-