Chapter 1.9
Multimodal and Federated Interaction

Frankie James
SAP Research, USA

Rama Gurram
SAP Research, USA

ABSTRACT

This chapter introduces the concepts of multimodal and federated interaction. Because multimodality means, simply, the combination of multiple modalities (or types of input and output), the authors first introduce some of the various modalities available for computer interaction. The chapter then discusses how multimodality can be used both in desktop and mobile computing environments. The goal of the chapter is to familiarize scholars and researchers with the range of topics covered under the heading "multimodality" and suggest new areas of research around the combination of modalities, as well as the combination of mobile and stationary computing devices to improve usability.

THE BASICS OF MULTIMODALITY

As was discussed in the introduction to "Liability," interaction with a computer (or computer-based device) can take place using a variety of different forms or modalities. On the input side, information can be transferred from a human operator to the computer via keyboards, keypads, touch screens, mice, joysticks, spoken language, or even gesture and motion sensors. Information can be output through visual displays (large and small), audio displays (including spoken text and non-speech sounds; see also "Mobile Speech Recognition"), and tactile displays (such as Braille or raised-line displays), as well as more exotic forms, such as force-feedback (haptic) joysticks and mice, and olfactory ("aroma-based") displays.

Each of the input and output channels mentioned above have their own benefits and limita-

tions. For example, large visual displays are able to present a great deal of information at the same time, which users can quickly scan to find the data that is relevant to their current needs. However, visual displays are not appropriate for blind users, and they are also inappropriate for mobile users who are not able to stand or sit in one place to read the display (or carry it with them while traveling).

The purpose of this chapter is to discuss the concepts of *multimodal* interaction, where two or more modalities (such as vision or audio) are combined. Multimodal interaction is frequently used to compensate for limitations in one interaction modality by providing a second one. For example, the limited visual display capabilities of a mobile device can be augmented by providing audio output, or speech input can be provided to a user with limited typing ability to increase data entry speed. Multiple modalities can also be combined within a single input (or output) to increase efficiency; a seminal example here is the combination of a spoken action command (e.g., "color this red") with mouse or touch-screen selection of the object to be acted upon. There are, of course, many other reasons for using multimodality, which will be discussed later in this chapter.

We will begin in this introduction with a discussion of the different forms of multimodality, as well as the different purposes to which multimodal interactions can be applied. In the next two sections, we address the use of multimodality for desktop applications and on mobile devices, where the reasons and methods for using multimodality can be quite different. We then discuss the concept of device federation, where multiple devices (each with their own available modalities) can be combined within a single interaction. Finally, we conclude with the chapter summary.

DIFFERENT FORMS

The basic definition of multimodality is the use of more than one modality within a single interface. The availability of both keyboard and voice input is one of the most common examples of multimodality, as is the use of both visual (text or graphical) and audio output. Most of the five classical human senses (sight, hearing, touch, smell, and taste) can be used for both the input and output sides. Each sense allows a broad range of possibilities.

Table 1 gives a brief list of the types of input and output that are associated with the senses.

The most common use of the sense of sight is in the visual presentation (output) of information on small and large displays. Sight can also be used for input: eye tracking can be used for selection or to gauge interest in a particular area of a screen, and retinal scanning can be used to identify the user.

Input options based on the sense of hearing include speech, for entering text or giving commands, speaker identification (to identify or authenticate the user), and even humming (Ghias et al., 1995). Audio output can be used for presenting written text (using text-to-speech), recorded audio files, document and interface structures, and sonifications of graph data [see the chapter "Mobile Speech Recognition" and James (1998) for an overview]. Speech input and audio-based output are useful in a variety of contexts, including mobile and vehicle-based scenarios, as well as accessibility.

The sense of touch is already commonly found in computer inputs today, through the use of keyboards, pointing devices, and touch screens. In addition to detecting simply that a key, button, or screen area has been clicked or touched, more advanced devices (such as game controllers and track pads) can also detect the amount of pressure exerted by the user. Handwriting and gesture are also gaining in popularity within certain contexts, along with the use of fingerprints for user iden-