CHAPTER 1

Definition and Types of Multimodal Interface

Multimodal interfaces support user input and processing of two or more modalities—such as speech, pen, touch and multi-touch, gestures, gaze, and virtual keyboard. These input modalities may coexist together on an interface, but be used either simultaneously or alternately. The input may involve recognition-based technologies (e.g., speech, gesture), simpler discrete input (e.g., keyboard, touch), or sensor-based information. Some of these modalities may be capable of expressing semantically rich information and creating new content (e.g., speech, writing, keyboard), while others are limited to making discrete selections and controlling the system display (e.g., touching a URL to open it, pinching to shrink a visual display). As will be discussed in this chapter, there are numerous types of multimodal interface with different characteristics that have proliferated during the past decade. This general trend toward multimodal interfaces aims to support more natural, flexible, and expressively powerful user input to computers, compared with past keyboard-and-mouse interfaces that are limited to discrete input.

This book focuses on interfaces involving multimodal user input capabilities, as distinct from the complementary and large body of literature on multimedia system output from a signal processing, circuits, networking, algorithms and software design viewpoint (Chen, 2015; Steinmetz, 2015; Steinmetz and Nahrstedt, 2004). It also is distinct from treatments of intelligent multimedia systems that emphasize presentation planning and application of artificial intelligence techniques to system processing and output (Maybury, 1993). Our treatment of multimodal input addresses the human-centered design of system capabilities available to users for accessing and interacting with computational devices, rather than the development of system output capabilities. Although it is common to do so, a multimodal interface may or may not be combined with multimedia system output to the user.

Throughout this book, we distinguish between keyboard-and-mouse interfaces that combine two relatively simple and discrete forms of input, from multimodal interfaces that expand the expressive power, flexibility, and naturalness of interfaces by incorporating recognition- and sensor-based information sources. In a multimodal interface, some modalities may involve active user input to a system (e.g., semantic content of speech, writing, or typing), while others process passively tracked continuous information that the user may not intend specifically as system input (e.g., visual processing of facial expressions, gaze location). Passively tracked information also can

include paralinguistic or low-level signal dimensions of input modalities like speech and writing, such as speech pitch or volume that can reveal energy level and emotional state.

This distinction between active input modes and *passive input modes* depends on whether the user is consciously deploying their actions with the intent of providing input to a computer system. Passive input modes and sensors can provide information to a multimodal system more unobtrusively and transparently to facilitate interactions, and their use is expanding rapidly. However, unless carefully designed, systems that process sensors and passive input modes can result in higher false alarms and be less intuitive to users. As will be discussed in Chapter 12, continuous monitoring of user behavior also risks violation of privacy, commercial exploitation, government surveillance, and other concerns that our society is struggling with today. See Table 1.1 for terms.

The most advanced multimodal interfaces are *fusion-based* ones that co-process two or more combined user input modes that can be entered simultaneously, such as speech and pen input, speech and lip movements, or speech and manual gesturing. They process meaning from two recognition-based technologies to fuse an overall interpretation based on joint meaning fragments from the two signal streams (see Chapter 8 for details). At the signal level, recognition technologies process continuous input streams (e.g., acoustic speech, digital ink trace) that contain high-bandwidth and often semantically rich information, which can be used for *content creation*—or to create and interact with a system's application content. Fusion-based multimodal interfaces have been developed and commercialized in recent years (see Chapters 7-9), and their interfaces have been designed and optimized to support a specific range of tasks. See Table 1.1 for italicized terminology.

At the other end of the spectrum, the simplest multimodal interfaces have input options that are used as *alternative modes*, rather than being co-processed. These rudimentary multimodal interfaces have been widely commercialized, for example on cell phones. They often offer more input options than a fusion-based multimodal system, such as speech, touch and multi-touch, gestures (e.g., pinching, flicking), stylus, keyboard,[1] and sensor-based controls (e.g., proximity for system activation, tilt for direction of screen display). In this regard, their interfaces typically involve a more "kitchen-sink" modality combination, which is less optimized for any specific set of modalities. These interfaces often include modes limited to *controlling the system display*, such as touching a URL to open a web page or pinching the screen to shrink its size. When input modes capable of

[1] New hybrid virtual keyboards on cell phones are becoming more flexible, multi-lingual, intelligent, and even multimodal. For example, some products such as Swype (Nuance, 2015a) support (1) four different interaction modes: tapping to type, gestural swiping of letters, handwriting letters, or spoken dictated letters; (2) choice of input language; (3) choice of level of recognition (stroke, character, sentence). Users can enable predictive input of subsequent letters. They also can enable data collection from certain applications, so the system learns and improves recognition. Some systems also support flexible shifting of languages during input, such as active use of "Hinglish" by Indian users who code mix. In this book, we distinguish between traditional tap-to-type keyboard interfaces, and these new hybrid virtual "keyboards" on mobile devices.

content creation are included in an alternative-mode multimodal interface (e.g., speech), recognizers may be used to process them but they are not processed jointly with another mode.

On mobile devices, some multimodal interfaces actually are *multimodal-multisensor* ones (see Table 1.1). These interfaces combine multimodal user input with one or more forms of sensor control involving contextual cues. For example, proximity of a cell phone to the user's ear may turn it on, or the angle of holding a phone may orient its display. When a user says or types, "Where are the nearest Italian restaurants?" sensor information about the user's location can constrain the listings and display them on a map. Users do not necessarily need to consciously engage sensor controls, although they may learn their contingencies over time and begin to do so. Multimodal-multisensor interfaces aim to transparently facilitate user interaction, especially while they are in mobile settings. This type of interface already has been commercialized widely on cell phones, wearables, in-vehicle interfaces, and other mobile devices.

On current multimodal interfaces, true alternative-mode interfaces actually are increasingly rare. For example, cell phone interfaces typically only process speech after a user presses to talk, which then engages speech processing. In other cases, speech input queries are not processed in isolation, but rather fused with sensor-based information (e.g., location) as illustrated in the above restaurant search example. In this respect, the simplest types of multimodal interface today typically use one mode or sensor to either engage or constrain a second modality.

Temporally cascaded multimodal interfaces process a combination of modalities that tend to be sequenced in a particular order, such as gaze directed at an object followed by speaking about it. When a temporally predictable sequence of modalities is present in users' behavior, the system can use partial information supplied by the earlier modality (e.g., gaze, touch) to constrain interpretation of information presented in the later-arriving mode (e.g., speech). Systems of this type potentially could process meaningful information trimodally, for example processing a user's gaze, then touch or pointing, followed by speech. Temporally cascaded multimodal interfaces potentially can powerfully constrain meaning interpretation, but they have been under-exploited.

As illustrated in the above example, some multimodal interfaces fuse information from more than two sources. However, this processing does not necessarily always assume a fixed temporal sequence. In multi-biometric interfaces, for which highly reliable person identification is imperative, three or more information sources processed jointly can achieve highly reliable identification that rules out imposters. Such systems emphasize fusing a larger number of information sources, uncorrelated information sources, and passively tracked behaviors or physical characteristics that are difficult to spoof (e.g., iris pattern, fingerprints, face recognition, gait, speech quality). To further improve reliable identification, this type of multimodal system often includes a multi-algorithm or parallel fusion approach (Ross and Poh, 2009). Commercial multi-biometric systems are described further in Chapter 9.

The long-term trend has been expansion in the number and type of information sources represented in multimodal interfaces, which has been especially noteworthy on current smart phones. This has included both active and passive recognition-based input modes. It also has involved rapid incorporation of new sensors, increasing the prevalence of multimodal-multisensor interfaces. One impact of these trends has been to multiply the strategies observed for effectively fusing information among sensors and modes in order to create new functionality. While new functionality often simply controls the interface display, increasingly it also is expanding the system's ability to interpret users' meaning more accurately. This may include processing that extends beyond handling individual input, to refining the accuracy of interpretation during interactive dialogues with follow-up queries. Figure 1.1 illustrates the progression of increasingly more powerful types of multimodal interface.

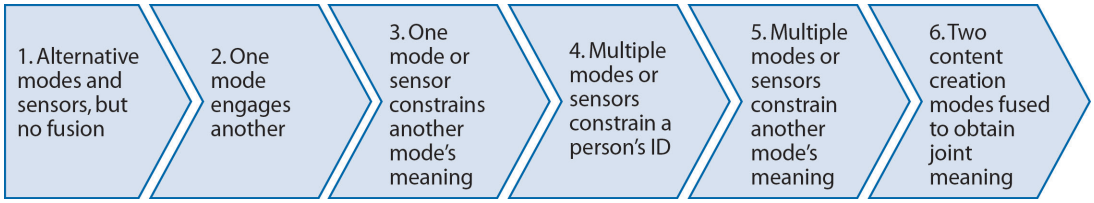

Figure 1.1: Different types of multimodal interface, which have been developed to support increasing complexity and expressive power, from: (1) alternative modes and sensors on the interface, but no fusion among them; (2) one modality used to engage another, such as touch to talk; (3) one mode or sensor constrains another's meaning interpretation, such as location while mobile determining what "nearby" means in a spoken query about restaurants; (4) multiple modes or sensors retrieve person identification from a database, as in multi-biometric systems; (5) multiple modes or sensors constrain another mode's meaning, such as GPS location, direction of movement, and pointing at a highway route to disambiguate the meaning of a spoken query about "gas stations up ahead"; and (6) two content creation modes are fused to produce a joint meaning interpretation, such as writing to encircle an area while speaking "show real estate." Beyond level 6 fusion systems, there are opportunities to increase expressive power in a number of ways. For example, interactive dialogue capabilities could be added to progressively constrain or correct information (see Section 2.3 examples). In addition, passively-tracked information (e.g., facial expressions, voice quality) could be added to provide further information about a user's attitude, emotional state, personality, and so forth, thereby resulting in a system with multiple content creation and passively tracked modes. From Oviatt (2012).

Table 1.1: Definition and types of multimodal interface
Multimodal Interfaces support input and processing of two or more modalities, such as speech, pen, touch and multi-touch, gestures, gaze, and virtual keyboard, which may be used simultaneously or alternately. User input modes can involve recognition-based technologies (e.g., speech) or discrete input (e.g., keyboard, touch). Some modes may express semantically rich information (e.g., pen, speech, keyboard), while others are limited to simple selection and manipulation actions that control the system display (e.g., gestures, touch, sensors).
Fusion-based Multimodal Interfaces co-process information from two or more input modalities. They aim to recognize naturally occurring forms of human language and behavior, which incorporate one or more recognition-based technologies (e.g., speech, pen, vision). More advanced ones process meaning from two modes involving recognition-based technologies, such as speech and gestures, to produce an overall meaning interpretation (see chapter 8). Simpler ones jointly process information from one mode or sensor (e.g., pointing/touching an object, location while mobile) to constrain the interpretation of another recognition-based mode (e.g., speech).
Alternative-Mode Multimodal Interfaces provide an interface with two or more input options, but users enter information with one modality at a time and the system processes each input individually.
Multimodal Interfaces for Content Creation incorporate high-bandwidth or semantically rich modes, which can be used to create, modify, and interact with system application content (e.g., drawing lake on a map). They typically involve a recognition-based technology.
Multimodal Interfaces for Controlling the System Display incorporate input modes limited to controlling the system or its display, such as zooming or turning the system on, rather than creating or interacting with application content. Examples include touch and multi-touch (e.g., for selection), gestures (e.g., flicking to paginate), and sensor-based controls (e.g., tilt for angle of screen display).
Active Input Modes are ones that are deployed by the user intentionally as explicit input to a computer system (e.g., speaking, writing, typing, gesturing, pointing).
Passive Input Modes refer to naturally occurring user behavior or actions that are recognized and processed by the system (e.g., facial expressions, gaze, physiological or brain wave patterns, sensor input such as location). They involve user or contextual input that is unobtrusively and passively monitored, without requiring any explicit user command to a computer.

Temporally Cascaded Multimodal Interfaces are ones that process two or more user modalities that tend to be sequenced in a particular temporal order (e.g., gaze, gesture, speech), such that partial information supplied by recognition of an earlier mode (e.g., gaze) constrains interpretation of a later one (e.g., manual selection), which then may jointly constrain interpretation of a third mode (e.g., speech). Such interfaces may combine active input modes, passive ones, or blend both types of input.

Multimodal-Multisensor Interfaces combine one or more user input modalities with sensor information that involves passive input from contextual cues (e.g., location, acceleration, proximity, tilt) that a user does not need to consciously engage. They aim to incorporate passively-tracked sensor input to transparently facilitate user interaction, which may be combined with an active input mode (e.g., speech) or a passive one (e.g., facial expression). The type and number of sensors incorporated into multimodal interfaces has been expanding rapidly on cell phones, in cars, robots, and other applications—resulting in explosive growth of multimodal-multisensor interfaces (see Chapter 9).

Visemes refer to the classification of visible lip movements that correspond with audible phonemes during continuous articulated speech. Many audio-visual speech recognition systems co-process these two sources of information multimodally to enhance the robustness of recognition.