Multimodality in Mobile Applications and Services

Maria Chiara Caschera
Istituto di Ricerche sulla Popolazione e le Politiche Sociali – CNR, Italy

Fernando Ferri
Istituto di Ricerche sulla Popolazione e le Politiche Sociali – CNR, Italy

Patrizia Grifoni
Istituto di Ricerche sulla Popolazione e le Politiche Sociali – CNR, Italy

INTRODUCTION

Multimediality and multimodality are concepts with multiple meanings. In Van den Anker and Arnold (1997), multimediality is defined as a way to present and convey information using several different media. Multimodality provides the user with multiple modalities of interacting with a system, beyond the traditional keyboard and mouse.

Both multimediality and multimodality refer to more than one communication channel. The diffusion of mobile devices and the development of their services and applications is connected with the natural communication approach preferred by users, which combines several modalities (speech, sketch, etc.) in order to communicate. It is therefore necessary to integrate the diverse modalities so that the features of the mobile device are more similar to the paradigm of human communication, making it simpler to use.

W3C (World Wide Web Consortium) activities have solved various mobile Web problems affecting the diffusion of mobile devices, such as practice Web navigation from mobile devices, multimodal interaction for the mobile Web, and multimedia and graphics for multimedia messaging. Diverse W3C working groups are involved in the discussion about device independence, multimodal Web access, and type of content for multimedia messaging. Characteristics identified by W3C according to the power and extensibility of XML (eXtensible Markup Language) enable the exchange of rich multimedia content and promote inter-user communication.

The greater opportunity to access and exchange information according to the content and opportunities offered by mobile devices are the two main elements of the growing interest in added-value services in different social environments. In fact, the opportunities offered by multimedia and multimodal technologies are important for any type of communication services, and as these technologies allow new ways of communicating, particular attention can be devoted to people with no technological ability or those with disabilities. This article introduces and discusses the problems and future scenarios and prospects of multimodality and mobile communication, analyzing the multimodal dialogue systems.

BACKGROUND

Multimodal applications combine visual information (involving images, text, sketches, and so on) with voice, gestures, and other modalities to provide powerful mobile applications, giving users the flexibility to choose one or more of the multiple interaction modalities. These systems break down the barriers in adopting mobile devices for added-value services.

Mobile applications and services generally involve both multimodality and multimediality, using different modalities and (as specified below) different communication channels. Two typical cases of mobile multimedia use are person-to-person communication and person-to-content communication (Ericsson, 2004). In addition, several applications require multimedia data, and numerous studies have been devoted to developing new ideas for methodologies, technologies, and algorithms for indexing, retrieving, compressing, transmitting, and integrating different types of data (Pham & Wong, 2004).

The use of multimodality and multimediality in mobile devices allows a simple, intuitive communication approach and generates new, richer services for users (Colby, 2002). When developing multimodal services, it is essential to consider perceptual speech, audio, and video quality for optimum communication system design and effective transmission planning and management in order to satisfy customer requirements (Kitawaki, 2004).

The following parameters should be considered when characterizing quality in advanced mobile devices: (1) wideband speech, audio, and video for multimedia; (2) noise reduction; and (3) speech recognition-synthesis for hands-free communication.

Multimedia and multimodal systems use different channels of communication, and specific devices are developed to increase the information flow between user and system. Nigay and Coutaz (1993) distinguish between the two, observing that a multimodal system is able to automatically model information content through a high level of abstraction. This difference leads to the definition of two main characteristics of multimodal interfaces:

- fusion among different data types and different input/output devices; and
- temporal constraints, imposed by information processing to and from input/output devices.

A multimodal system is an hw/sw system that allows one to receive, to interpret, and to process input, and that generates as output two or more interactive modalities in an integrated and coordinated way.

Communication among people is often multimodal, and it is obtained combining different modalities. Multimodal interfaces allow several modalities of communication to be harmoniously integrated, making the system's communication characteristics more similar to the human approach.

Multimodal interfaces provide the user with multiple interaction paradigms through different types of communication input. Data fusion is one of the main problems in human-computer interaction, where each datum is generated through a distinct interaction mode. Furthermore, the management of these multiple processes includes synchronization and selection of the predominant mode. Consequently, an important issue in multimodal interaction is the integration and synchronization of several modalities in a single system. In literature two approaches are often used: signal fusion, and information fusion at the semantic level.

The first approach is preferred for matching and synchronizing modalities as speech and labial movement. The semantic fusion is used for modalities that differ in a temporal scale. In this approach, time is very important because chunks of information with different modalities are considered, and integrated if they are temporally close. The integration can be carried out using an intermediate approach between the signal integration and the semantic fusion.

The relation among modal components can be classified as follows (Bellik, 2001):

- **Active:** (Act in following tables)—when two events, produced by two different devices, cannot be completely and correctly interpreted without ambiguities if one of the two events is unknown.
- **Passive:** (Pas in following tables)—when an event produced by a given device cannot be completely and correctly interpreted without ambiguities if the state of the other devices is unknown.

The input synchronization of a multimodal system can be defined as:

- **Sequential:** (Seq in following tables)—if the interpretation of the interactive step depends on one mode and the modalities can be considered one by one.
- **Time-Independent Synchronized:** (TIS in following tables)—if the interpretation of the interactive step depends on two or more modalities and the modes are simultaneous.
- **Time-Dependent Synchronized:** (TDS in following tables)—if the interpretation of the interactive step depends on two or more modalities and the semantic dependence of the modalities has a close temporal relationship.

There are several levels of synchronization (W3C):

- **Event-Level:** If the inputs of one mode are received as events and immediately propagated to another mode.
- **Field-Level:** If the inputs of one mode are propagated to another mode after a user has changed the input field or the interaction with a field is terminated.
- **Form-Level:** If the inputs of one mode are propagated to another mode after a particular point of the interaction has been achieved.
- **Session-Level:** If the inputs of one mode are propagated to another mode after an explicit changeover of mode.

The semantic fusion of modal input occurs in two steps: (1) the first matches the modalities to obtain a low-level interpretation module, by grouping the input events in multimodal events; and (2) the second transfers the multimodal inputs to the high-level interpretation module, in order to obtain the meaning of their events. This high-level interpretation defines the type of actions that will be triggered by the user and the used parameters. These parameterized actions are passed to the application dialog manager to start their execution.

MAIN FOCUS OF THE ARTICLE

There is an emerging need for integration among the various input modalities, through signal integration and semantic fusion, and an additional need to disambiguate the various input modalities and coordinate output modalities, to enable the user to have a range of integrated, coordinated interaction modalities.

Martin and Toward (1997) proposed a theoretical framework for studying and designing multimodal systems based on a classification of six basic types of cooperation between modalities: