Chapter II
Analyzing Multimodal Interaction

Fernando Ferri
Istituto di Ricerche sulla Popolazione e le Politiche Sociali
Consiglio Nazionale delle Ricerche, Italy

Stefano Paolozzi
Istituto di Ricerche sulla Popolazione e le Politiche Sociali
Consiglio Nazionale delle Ricerche, Italy

ABSTRACT

Human-to-human conversation remains such a significant part of our working activities because of its naturalness. Multimodal interaction systems combine visual information with voice, gestures, and other modalities to provide flexible and powerful dialogue approaches. The use of integrated multiple input modes enables users to benefit from the natural approach used in human communication. In this paper, after introducing some definitions and concepts related to multimodal environment, we describe the different approaches used in the multimodal interaction fields showing both theoretical research and multimodal systems implementation. In particular, we will address those approaches that use the new semantic technologies as well as the ones related to the Semantic Web fields.

INTRODUCTION AND BACKGROUND

There is a great potential for combining speech and gestures and other “modalities” to improve human-computer interaction because this kind of communication resembles more and more the natural communication humans use every day with each other.

Nowadays, there is an increasing demand for a human-centred system architecture with which humans can naturally interact so that they do no longer have to adapt to the computers, but vice versa. Therefore, it is important that the user can interact with the system in the same way as with other humans, via different modalities such as speech, sketch, gestures, etc. This kind of multimodal human-machine interaction facilitates the
Analyzing Multimodal Interaction

communication for the user of course, whereas it is quite challenging from the system’s point of view.

For example, we have to cope with spontaneous speech and gestures, bad acoustical and visual conditions, different dialects and different light conditions in a room and even ungrammatical or elliptical utterances which still have to be understood correctly by the system. Therefore, we need a multimodal interaction where missing or wrongly recognized information could be resolved by adding information from other knowledge sources.

The advantages of multimodal interaction are evident if we consider practical examples. A typical example for multimodal man machine interaction which involve speech and gestures is the "Put That There" from Bolt (Bolt, 1980). Since that time, lots of research has been done in the area of speech recognition and dialogue management so that we are now in the position to integrate continuous speech and to have a more natural interaction. Although the technology was much worse in these times, the vision was very similar: to build an integrated multimodal architecture which fulfils the human needs. The two modalities can complement each other easily so that ambiguities can be resolved by sensor fusion. This complementarity has already been evaluated by different researchers and the results showed that users are able to work with a multimodal system in a more robust and stable way than with a unimodal one. The analysis of the input of each modality could therefore serve for mutual disambiguation. For example, gestures can easily complement to the pure speech input for anaphora resolution.

Another reason for multimodal interaction is the fact that in some cases the verbal description of a specific concept is too long or too complicated compared to the corresponding gesture (or even a sketch) and in these cases humans tend to prefer deictic gestures (or simple sketches) than spoken words. On the other hand, considering for example the interaction of speech and gesture modalities, there are some cases, where, for example deictic gestures are not used because the object in question is too small, it is too far away from the user, it belongs to a group of objects, etc.; here, also the principles of Gestalt theory have to be taken into account which determine whether somebody pointed to a single object or to a group of objects.

Moreover, there it has been also empirically demonstrated that the user performance is better in multimodal systems than in unimodal ones, as explained in several works (Oviatt, 1999; Cohen et al., 1997; Oviatt et al., 2004). Of course, it is clear that the importance to have a multimodal system than a unimodal one strictly depends on the type of action being performed by the user. For instance as mentioned by Oviatt (Oviatt, 1999), gesture-based inputs are advantageous, whenever spatial tasks have to be done. Although there are no actual spatial tasks in our case, there are some situations where the verbal description is much more difficult than a gesture and in these cases, users may prefer gestures.

As shown by several studies, speech seems to be the more important modality which is supported by gestures as in natural human-human communication (Corradini et al., 2002) This means that the spoken language guides the interpretation of the gesture; for example, the use of demonstrative pronouns indicates the possible appearance of a gesture. Therefore, several studies have been proved that speech and gestures modalities are co-expressive (Quek et al., 2002; McNeill & Duncan, 2000) which means that they present the same semantic concept, although different modalities are used. This observation can be extended also to other modalities that may interact with speech modality.

The advantage of using multimodal systems instead of adopting unimodal ones relies not only in the fact that there are more efficient, but also the number of critical errors can be reduced. In other words, the multimodal systems are more flexible