Chapter XVI
Speech Driven Interaction in Mobile Multimodality

Giovanni Frattini
Engineering IT, Italy

Francesco Gaudino
Engineering IT, Italy

Vladimiro Scotto di Carlo
Engineering IT, Italy

Fabio Corvino
Engineering IT, Italy

Pierpaolo Petriccione
Engineering IT, Italy

Gianluca Supino
Engineering IT, Italy

ABSTRACT

This chapter introduces a possible architecture for building mobile multimodal applications and our experiences in this domain. Mobile devices are becoming increasingly powerful and sophisticated and their use more and more diffused; the need for new and complex services has consequently amplified. New generation mobile applications must be able to manage convenient inputs and outputs modes to make the dialog natural, intuitive, and user centric. Simultaneous use of complementary communication channels (multimodality), such as voice, keyboard, stylus, leads to a more complicated input processing system but it’s a way to simplify the interaction. A speech driven interaction between user and service delivery systems may be the ideal solution for the development of ubiquitous and context aware applications: besides being a very instinctive and convenient way to express complex questions, speech is also the best option when eyes and hands are busy.

INTRODUCTION

In this chapter we report some of the experiences we have gained working on a new mobile multimodal speech driven software solution, in a research project called CHAT (Frattini et al. 2006).

More specifically, we have concentrated efforts on mobile multimodal services for supporting cultural heritage fruition (museums, archeological parks, etc.) and e-learning. Multimodal services can be synergic or alternate. Synergic multimodality, which refers to the simultaneous use of different modalities (speech, sketch, handwritten,
Speech Driven Interaction in Mobile Multimodality

keyboard) in a single interaction act, could introduce real benefits to mobile users when keyboard-based interaction is difficult. On the other hand, alternate multimodality, which is characterized by a sequence of unimodal messages, has been investigated during past years (Frattini et al. 2006), without an evident commercial success.

A platform for synergic mobile multimodal services is a complex system. Some of the issues discussed in this chapter are well known, others are related to users mobility. We will consider:

- The logic for establishing the user intention and the task to be executed on the server side.
- The target terminals and the client software environment for enabling an optimal user experience.
- The underlying technological infrastructures in terms of network protocols and efficiency.

Synergic multimodality encompasses different processes: one of them, called “fusion”, must combine the different modes in order to match the user intention. We can affirm that a software architecture for building multimodal services must have a fusion module (Frattini et al. 2006). Our choice has been to implement a speech-driven fusion process: the role of voice inputs is more relevant than other possible input modalities. Voice is probably the most appropriate and instinctive mean to express complex commands, especially when the information content of user requests becomes more and more rich and complicated. Nevertheless, using complex modalities, like voice, has some important consequences: speech recognition must be as accurate as possible. Furthermore, not all the commercial handsets are able to host a speech recogniser and process vocal inputs locally. Thus, distributing recognition processes can help to improve recognition quality. In particular, as the information content of user requests becomes more and more complicated, the ideal solution is to acquire and transport vocal signals over a mobile network and, thus processing them on more powerful hardware. Once on the server, different natural language understanding algorithms can be applied. The meaning of linguistic utterances can be derived using formal structures, or meaning representations. The need for these representation models arises while trying to bridge the gap between linguistic inputs and non-linguistic knowledge of the world needed to perform tasks involving the meaning of linguistic inputs. Multimodality can help in formulating more appropriate hypothesis about the context in which the voice command must be placed and could help in improving the machine understanding process.

Simply accessing the phonological, morphological, and syntactic representations of sentences may not be enough in order to accomplish a task. For example, answering questions requires background knowledge about the topic of the question, about the way questions are usually asked and how such questions are usually answered. Therefore the use of domain-specific knowledge is required to correctly interpret natural language inputs.

An empirical approach to constructing natural language processing systems starts from a training corpus comprising sentences paired with appropriate translations into formal queries. Learning algorithms are utilized to analyse the training data and produce a semantic parser that can map subsequent input sentences into appropriate tasks. Some techniques attempt to extend statistical approaches, which have been successful in domains such as speech recognition, to the semantic parsing problem.

In our proposal, the process of semantic inputs disambiguation is based on the collaboration of a stochastic classifier and a rule based engine. A domain specific data corpus has been manually labelled to build the model language while a named entity recogniser works as classifier, in order to extract meaningful semantic contents. This approach allows ignoring false starts or hesitations.