Towards Multimodal: A Telecom Perspective

Keith Waters
France Telecom Research and Development

June 16th 2004

Abstract

The emerging area of Multimodality can be well demonstrated in the operation of a Telecom. Within France Telecom there are thriving fix-line, Internet and mobile businesses that can benefit from the use of Multimodal interactions. In todays networks, Multimodal interactions are particularly relevant to the mobile space, where constraints of the network and handset have a driving influence on the interactions available to everyday mobile users. This paper describes some of the concurrent technical issues involved in building Multimodal systems within a Telecom and provides indicators to the future Multimodal concepts.

1 Introduction

Access to rich data sources, via a Telecom's network, presents interesting challenges for general purpose Multimodality. For instance, the constraints of mobile handsets with limited input and output functionality, dictate a particular interaction style. Often mobile devices have only a couple of modes available, such as speech, keypad or pen of which one may be unreliable. Therefore mode redundancy is perceived as one of the first commercial efficiencies of a Multimodal system where fall-back modes can remain operational. Within a fixed environment Multimodal solutions have fewer constraints with a wider range of available input and output modalities and much higher data bandwidth. For example, in both the home and office multiple human and environmental sensors can be blended into one seamless interaction. For Multimodal Internet applications access to local and remote resources are critical, however such resources may well have long response latencies. In each case – fixed, Internet and mobile – Multimodal solutions present some emerging general purpose requirements.

To date Multimodality activities typically consider the primary human modes of interaction via voice [4], touch (pens/DTMF), passive computer vision sensors [3] and feedback through haptic interfaces [5]. Clearly there is a desire to construct Multimodal systems that closely mimic and leverage human interaction styles reminiscent of Bolts *"Put-that There"* [2]. Consequently, research has placed emphasis on fusing ambiguous sensor data, as well as constructing improved sensors for human-like interactions involving gesture and gaze models,

improved speech recognition and even combinations of audio-visual speech perception [1]. While academic institutions report progress, commercial reliability remains low.

As research progress continutes, there is an emerging alternative that presents itself in the same Multimodal context, namely interactions that are determined by devices and their environmental situation. For example, sensors allow devices to communicate their status, configuration and behavior within a Multimodal system. Such sensors define machine-based Multimodality. A useful example to consider is Location Based Services (LBS) which provides information about the devices position in both time and space. Such sensors are early examples of a discrete unambiguous modal sensor. Other temporal and spatial environmental sensors abound; orientation, temperature, light, signal strength, battery strength etc., provide valuable contextual information that can shape a users interaction. Such machine sensors remove much of the ambiguty presented by human-like sensors making it straight-forward to envision Multimodal interactions directly shaped by a devices' context and environment settings. A few key advantages are presented by this approach: firstly the lack of algorithmic complexity required to synthesize results based on modal fusion and secondly, a far simpler integration schemas can be constructed to generate outputs.

2 Modal sensors: the rise of the machine

Within the concept of a sensor it is important to consider how they become active, generate output and communicate within an application framework. Key technical concepts of how the data the sensor is collecting can be accessed presents unique challenges to the application builder. For example, does the application push or pull data from the sensor? Is the behavior of the sensor blocking or non-blocking? What are the concepts of persistence and state? How does a sensor notify the application that it is disconnected? All of these concepts are tractable technical issues that present unique interfaces to a Multimodal system. The W3C System and Environment Framework is an example of how properties and attributes can communicate via DOM level 2 events information to a Multimodal application developer [6].

2.1 Example: Network Signal Strength

To illustate this paper's position, a straight-forward example of a machine-based Multimodal sensor that dynamically modifies its characterisitics overtime is provided. In a mobile context it is useful to determine the networks signal strength and how it may vary over time and space. In such an application notification that the user is about to traverse a networks *blackhole event horizon* can trigger a variety of Multimodal application behaviors.

3 Conclusions

This paper suggests that a simpler approach to Multimodality can be created by dealing with a device's environmental conditions and sensing capabilities. Device-based Multimodal integration can remove much of the sensor ambiguity issues on which many Multimodal systems are currently constructed. In doing so, the complexity of modal fusion and integration can be dramatically reduced to a simplified model of event responses.

References

[1] C. Benoit B. Le Goff. Audio-visual speech synthesis from French text: Eight years of models, design and evaluation at the ICP. *Speech Communication*, 26:117–129, 1998.

[2] R. A. Bolt. Put-that-there: Voice and gesture at the graphics interface. *Computer Graphics*, 14(3):262–270, Yuly 1980.

[3] J. Woodfill M. Harville G. Gordon T. Darrell. Integrated person tracking using stereo, color and pattern detection. *International Journal of Computer Vision*, 37(2):175–185, 2000.

[4] Voice Browser Working Group. Voice Extensible Markup Language (VoiceXML) 2.1 http://www.w3c.org/2004/wd-voicexml21-200403, March 2004.

[5] K. Hirota and M. Hirose. Providing force feedback in virtual environments. *Computer Graphics*, 15(5):22–30, Sept 1995.

[6] W3C. Multimodal interaction activity: http://www.w3c.org/2002/mmi/ System and Environment.