Talking Heads for the Web: What for?

Koray Balci, Fabio Pianesi and Massimo Zancanaro
ITC-irst

Introduction

Talking heads and, more generally, human-like embodied interfaces have become popular as the front end to many web sites and as part of many computer applications. In most cases, however, these interfaces have just the graphical appearance of a body without any internal representation of the communicative functions of the latter.

Human-like interfaces are another argument in the long-term debate on anthropomorphism of the interface (see for example, Shneidermann, 1998 and Reeves and Nass,1996). While substantial evidence is available that resorting to well designed human-like interfaces able to properly model communicative functions may improve usability of computer systems (see among others Cassell et al. 2000 and Rickenberg and Reeves, 2000), especially in the areas of tutoring systems or edutainment, it is still not clear how well this interaction paradigm can fit into the hypermedia-based interaction of the World Wide Web.

While it is likely that on the desktop computer the navigation metaphor will be hard to die, one can imagine that on new devices such as web-enabled phones, e-books and perhaps even iTV, the conversational metaphor laying behind human-like interfaces might have a chance. In our own work, we experimented with a museum mobile guide using virtual characters to encode some dimensions of user adaptation to help the user seamlessly manage the transitions between the mobile device and the kiosks during interaction (Rocchi et al. 2004). When adaptivity enters into play; that is when the system is able to modify the information space to tailor it to the specific interests or knowledge of the user (Brusilosvky, 2003), the navigation metaphor could turn out to be confusing. Since it is based on a spatial metaphor, where the pages of the hypermedia are like places where the user can go, be at, or return to, the adaptation process may change the topological configurations, thus, disorienting the user (Dale et al., 1997). In either situation, even when shifting to a conversational metaphor, the WWW infrastructure and its huge information space may be effectively employed.

There are at least three relevant issues to consider for this goal. The first issue concerns the need for standard ways of representing both the low-level animation primitives and the high-level communication directives, both being compatible with the WWW infrastructure. Indeed, for both levels there are standard approaches and languages even if their use, especially in commercial products, is not widespread. For the low-level, MPEG-4 Facial Animation describes the steps to create a talking agent by defining various parameters (Pandzic and Forchheimer, 2002). For the high-level, there are many different proposals; among them, APML (De Carolis et al. 2002), an XML-based language to specific performatives and emotions, is a promising one.

Another issue is the availability of tools for editing and deploying different faces (or bodies). Furthermore, in order to better understand the effectiveness of human-like interfaces, experimentation and evaluation with real users is mandatory. An important

aspect that must also be addressed is the role that emotions can play at the interface, with a special emphasis on cross-cultural difference.
In this position paper, we briefly introduce a tool called Xface (Balci, 2004), an open source, platform independent toolkit for developing 3D talking agents based on MPEG-4 FA (Facial Animation) standard and we briefly summarize some initial studies about cross-cultural recognition of emotions in synthetic faces.

Xface Toolkit

Xface provides a set of tools for generating 3D talking agents. The target audience is both researchers working on similar topics and developers in the software industry.
Xface is being developed using C++ programming language and incorporating object oriented techniques. Because of the wide audience we aim at, the architecture of Xface is meant to be configurable and easy to extend. All the pieces in the toolkit are operating system independent, and can be compiled with any ANSI C++ standard compliant compiler. For the time being rendering relies on OpenGL API (Application Programming Interface). Modular architecture makes the support of other rendering APIs almost transparent to application developers. The library is optimized enough to achieve satisfactory frame rates (minimum 25 frames per second are required for FAP generating tool) with high polygon count (12000 polygons) using modest hardware.
Xface is based on MPEG-4 FA specifications as previously explained. For the generation of MPEG-4 FAP streams, Xface relies on apml2fap tool (Lavagetto and Pockaj), which parses APML scripts and generates FAPs. APML provides us with a simple way to define emotions and create the animation parameters. For speech synthesis, we use another open source tool, Festival (Black and Taylor, 1997).
The current state of the toolkit involves three pieces of software as output of the Xface project. These are the core library, an editor for preparation of faces, and a sample player.
Xface is available as Open Source distribution from http://xface.itc.it.
In the next release, Xface will also support X3D file format which is a recommendation by W3C for web content management in the near future. (http://www.web3d.org/x3d/).

Preliminary Cross-cultural Evaluation of Expressiveness in Synthetic Faces

Xface has been developing in the context of the European-funded project PF-Star (http://pf-star.itc.it), within the same project we are also conducting preliminary cross-evaluation experiments with the double aim of evaluating the possibility of exchanging FAP data between the involved sites and assessing the adequacy of the emotional facial gestures performed by talking heads (Beskow et al. 2004). The results provide initial insights to the way people from various cultures react to natural and synthetic facial expressions produced in different cultural settings and to the potentials and limits of FAP data exchange.
One group of Italian (47 volunteer students from the University of Trieste) and one group of Swedish participants (30 volunteers from the University of Stockholm and KTH) were confronted with four blocks of 12 video-files each: 1) Italian actor, 2) Swedish actor, 3) Swedish synthetic face playing both Italian and Swedish FAP-files, and 4) Italian synthetic face playing both Italian and Swedish FAP-files, for a total of 48 stimuli per participant. Two nonsense words, ABBA and ADDA, uttered with three emotional states

(happy, angry and neutral), were selected from the common sub-set of data. The stimuli were played without the audio. The results of this preliminary evaluation show that there are no differences in the way participants, from either cultural settings (Italian vs. Swedish), react to natural and synthetic facial expressions produced in different cultural settings. Differences did, however, emerge as to the provenance of FAPs. Drawing clearer conclusions is not possible at this point since many factors related to cross-sites differences in recording conditions may have affected the results. We are planning more experiments and observations in order to address these issues.

References

Balci, K. Xface: MPEG-4 based open source toolkit for 3d facial animation. In Proc. Advance Visual Interfaces, 2004

Beskow J., Cerrato L., Cosi P., Costantini E., Nordstrand M., Pianesi F., Prete M., Svanfeldt G. Preliminary Cross-cultural Evaluation of Expressiveness in Synthetic Faces. In Proceedings of ADS2004.

Black A. W. and Taylor P. A. The Festival Speech Synthesis System: System Documentation, HCRC/TR-83, v1.1, 1997.

Brusilovsky, P. Adaptive hypermedia. *User Modeling and User Adapted Interaction, Ten Year Anniversary Issue (Alfred Kobsa, ed.) 11 (1/2)*, 87-110, 2001

Cassell J., Bickmore T. , Vilhjálmsson H. and Yan H. (2000) More than just a pretty face: affordances of embodiment. In Proceedings of the 5th international conference on Intelligent user interfaces, New Orleans.

Dale, R., Milosavljevic, M., and Oberlander, J. The Web as dialogue: the role of natural language generation in hypertext. Presented at the AAAI Spring Symposium on Natural Language Processing and the Web. Stanford, Ca., March 1997

De Carolis, B., V. Carofiglio, M. Bilvi & C. Pelachaud (2002). 'APML, a Mark-up Language for Believable Behavior Generation'. In: Proc. of AAMAS Workshop Embodied Conversational Agents: Let's Specify and Compare Them!, Bologna, Italy, July 2002

Lavagetto F. and Pockaj R. The Facial Animation Engine: Towards a High-Level Interface for Design of MPEG-4 Compliant Animated Faces. In IEEE Transaction on Circuits and Systems for Video Technology, 9(2):277–289, 1999.

Pandzic I. and Forchheimer R. Animation: The Standard, Implementation Applications. Wiley, 2002.

Reeves B, and Nass C. (1996) *The media equation: how people treat computers, television, and new media like real people and places*. Cambridge University Press

Rickenberg R., Reeves B. (2000) The effects of animated characters on anxiety, task performance, and evaluations of user interfaces. In Proceedings of Proceedings of CHI 2000.

Rocchi C., Stock O., Zancanaro M., Kruppa M., and Krüger A. The Museum Visit: Generating Seamless Personalized Presentations on Multiple Devices. In Proceedings of Intelligent User Interfaces 2004. Madeira, January 2004.

Shneiderman, B. (1998). *Designing the user interface. Strategies for effective human-computer interaction*. Reading, MA: Addison-Wesley