A webGL talking head for mobile devices

Alberto Benin
ISTC-CNR
Via Martiri della libertà 2
35137 Padova, Italy
0039 498271819
alberto.benin@pd.istc.cnr.it

Piero Cosi
ISTC-CNR
Via Martiri della libertà 2
35137 Padova, Italy
0039 498271822
piero.cosi@pd.istc.cnr.it

G.Riccardo Leone
ISTC-CNR
Via Martiri della libertà 2
35137 Padova, Italy
0039 498271822
riccardo.leone@pd.istc.cnr.it

ABSTRACT
Luciaweb is a 3D Italian talking avatar based on the new WebGL technology. WebGL is the standard programming library to develop 3D computer graphics inside the web browsers. In the last year we developed a facial animation system based on this library to interact with the user in a bimodal way. The overall system is a client-server application using the http protocol: we have a client(a browser or an app) and a web server. No software download and no plugin are required. All the software reside on the server and the visualization player is delivered inside the html pages that the client ask at the beginning of the connection. On the server side a software called AudioVideo Engine generates the phonemes and visemes information needed for the animation. The demo called Emotional Parrot shows the ability to reproduce the same input in different emotional states. This is the first WebGL software running on iOS device ever.

Categories and Subject Descriptors
I.3.7 [**Computer Graphics**]: - Three dimentional graphic and realism – *animation, color, shading, shadowing and texture, virtual reality.*.

General Terms
Design, Experimentation, Human Factors, Languages

Keywords
WebGL, talking head, MPEG-4, mobile, iOS, speech synthesis

1. INTRODUCTION
Face to face communication is the main element of human-human interaction because both acoustic and visual signal simultaneously convey linguistic, extra linguistic and paralinguistic information. Therefore facial animation is a research topic since the early 70's and many different principles, models and animations have been proposed over years [1, 2]. An efficient coding of shape and animation of human face was included in the MPEG-4 international standard [3]. At ISTC-CNR of Padua we developed LUCIA talking head, an open source facial animation framework [4]. With the introduction of WebGL [5], which is 3D graphics for web browsers, we enhanced the possibility for Lucia to be embedded in any internet site [6]. Now it's time for mobile devices. As far as we know this is the first webGL native application in the world running on iOS mobile devices.

Figure 1: Luciaweb on iOS device: text input during the Parrot Mode.

2. SYSTEM'S ARCHITECTURE
Luciaweb follows the common client-server paradigm. First off the client (a web browser or a mobile device application) opens a connection with the server; the answer is an HTML5 web-page which delivers the multimedia contents to start the MPEG4 player.

2.1 The webGL client
The typical WebGL application is composed by three parts: the standard html code, the main JavaScript program and a new shading language section. The html section is intended mainly for user interaction; the JavaScript part is the core of the application: the graphic library itself, all the matrix manipulation, support and utility functions take place here; the input from the user is connected with JavaScript variables via

ad-hoc event- driven procedures. The novelty is the third part which is the Shading Language code. This software runs on the Video Card. It is called GLSL and it derives from the C programming language. Actually these are the instructions that calculate every pixel color value on the screen whenever the drawing function is called in the JavaScript main program. To be able to change the values of the GLSL variables from the JavaScript WebGL Application Program Interface implements special methods to connect them with JavaScript objects, arrays and variables. During the initialization of the WebGL page the shader code is compiled and copied to the video card memory ready to be executed on the Graphic Processing Unit. At the beginning of the connection model parts data are fetched using the lightweight data-interchange format JSON[7]. This is the only moment where you could wait for a while because of the amount of the data to be transmitted, while right after this phase all the facial movements are almost real time.

2.2 The Audio Video Engine Server

Audio Video speech synthesis, that is the automatic generation of voice and facial animation parameters from arbitrary text, is based on parametric descriptions of both the acoustic and visual speech modalities. The acoustic speech synthesis uses an Italian version of the FESTIVAL di-phone TTS synthesizer [8] modified with emotive/expressive capabilities: the APML/VSML mark up language [9] for behavior specification permits to specify how to markup the verbal part of a dialog in order to modify the graphical and the speech parameters that an animated agent need to produce the required expressions. For the visual speech synthesis a data-driven procedure was utilized: visual data are physically extracted by an automatic opto-tracking movement analyzer for 3D kinematics data acquisition called ELITE [10]. The 3D data coordinates of some reflecting markers positioned on the actor face are recorded and collected, together with their velocity and acceleration, simultaneously with the co-produced speech. Using PRAAT [11], we obtain parameters that are quite significant in characterizing emotive/expressive speech [12]. In order to simplify and automates many of the operation needed for building-up the 3D avatar from the motion-captured data we developed INTERFACE [13], an integrated software designed and implemented in Matlab. To reproduce realistic facial animation in presence of co-articulation, a modified version of the Cohen-Massaro co-articulation model [14] has been adopted for LUCIA [15]

3. THE EMOTIONAL PARROT DEMO

To see the capabilities of emotional synthesis of our avatar you can use Lucia in “Parrot Mode”: she repeats any input text you enter and she can do it in six different emotional ways: joy, surprise, fear, anger, sadness, disgust. For this classification we take inspiration from [16] The demo is perfectly fluid on normal computer (about 60 fps) while it suffers of some frames skipping on low computational machine (7 fps on iPad 2).

4. CONCLUSION

Luciaweb is an MPEG-4 standard FAPs driven facial animation Italian talking head. It is a decoder compatible with the ”Predictable Facial Animation Object Profile”. It has a high quality 3D model and a fine co-articulation model, which is automatically trained by real data, used to animate the face. It runs on any WebGL compatible browser and now, first in the world, on iOS mobile devices (iPhone and iPad) It reproduces six different emotional states of the input text in “parrot mode”.

5. FUTURE WORK

In the near future we will integrate speech recognition to have double input channel and we will test the performance of the Mary TTS synthesis engine [17] for the Italian language. We will work on dynamic mesh reduction to enhance the frame rate on slow computational hardware.

6. REFERENCES

[1] F. I. Parke, F.I. and Waters, K. 1996. *Computer facial animation.* Natick, MA, USA: A. K. Peters, Ltd.

[2] Ekman, P. and Friesen, W. 1978. Facial action coding system. In *Consulting Psychologist*.

[3] Pandzic, I.S. and Forchheimer, R. Eds, 2003. *MPEG-4 Facial Animation: The Standard, Implementation and Applications*. John Wiley & Sons, Inc.

[4] Leone, G.R., Paci, G. and Cosi, P. 2011. LUCIA: An open source 3D expressive avatar for multimodal h.c.i. In *proc. of INTETAIN2011*

[5] WebGL, http://www.khronos.org/webgl/.

[6] Leone, G.R. and Cosi, P. 2011. Lucia-WebGL: a web based Italian MPEG-4 Talking Head. In *proc. of AVSP2011*

[7] JSON, http://www.json.org/.

[8] P. Cosi, P., Tesser, F, Gretter, R. and Avesani, C. 2001. Festival speaks italian! In *proc. of Eurospeech 2001*.

[9] Carolis, B.D., Pelachaud, C., Poggi, I. and Steedman, M., 2004. Apml, a mark-up language for believable behavior generation. *Life-Like Characters*, 65–85.

[10] Ferrigno, G. and Pedotti, A. 1985. Elite: A digital dedicated hardware system for movement analysis via real-time tv signal processing. *IEEE Trans. on Biomedical Eng.*

[11] Boersma, P. 1996. A system for doing phonetics by computer. *Glot International*,.341–345.

[12] Drioli, C., Cosi, P., Tesser, F. and Tisato, G. 2003. Emotions and voice quality: Experiments with sinusoidal modeling. In *proc. of Voqual 2003*. Geneva,127–132.

[13] Tisato, G., Drioli, C., Cosi, P. and Tesser, F. 2005. Interface: a new tool for building emotive/expressive talking heads. In *proc. of INTERSPEECH 2005*.

[14] Cosi, P. and Perin, G. 2002. Labial co-articulation modeling for realistic facial animation. In *proc. of ICMI 2002*.

[15] Cosi, P., Fusaro, A. and Tisato, G. 2003. Lucia a new italian talking- head based on a modified cohen-massaro labial co-articulation model. In *proc. of Eurospeech 2003*.

[16] Ruttkay, Z., Noot, H. and Hagen, P. 2003. Emotion Disc and Emotion Squares: tools to explore the facial expression space. *Computer Graphics Forum*, 22(1), 49-53.

[17] M. Schröder, M. and Trouvain, J. 2003. The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *Int. J. of Speech Technology*