A FACIAL ANIMATION FRAMEWORK WITH EMOTIVE/EXPRESSIVE CAPABILITIES

G. Riccardo Leone and Piero Cosi
Institute of Cognitive Sciences and Technologies – National Researh Council
Via Martiri della libertà 2, 35137 Padova, Italy

ABSTRACT

LUCIA is an MPEG-4 facial animation system developed at ISTC-CNR·. It works on standard Facial Animation Parameters and speaks with the Italian version of FESTIVAL TTS. To achieve an emotive/expressive talking head LUCIA was build from real human data physically extracted by ELITE optotracking movement analyzer. LUCIA can copy a real human by reproducing the movements of passive markers positioned on his face and recorded by the ELITE device or can be driven by an emotional XML tagged input text, thus realizing a true audio/visual emotive/expressive synthesis. Synchronization between visual and audio data is very important in order to create the correct WAV and FAP files needed for the animation. LUCIA's voice is based on the ISTC Italian version of FESTIVAL-MBROLA packages, modified by means of an appropriate APML/VSML tagged language. LUCIA is available in two different versions: an open source framework and the "work in progress" WebGL

KEYWORDS

Talking head, facial animation, mpeg4, affective computing, TTS, WebGL

1. INTRODUCTION

There are many ways to control a synthetic talking face. Among them, geometric parameterization [Massaro et al., 2000], morphing between target speech shapes [Bregler.et al., 1997], muscle and pseudo-muscle models [Terzopoulos et al., 1995 - Vatikiotis-Bateson et al.,1996], appear the most attractive.

Growing interest have encountered text to audiovisual systems [Beskow, J., 1995 - LeGoff, B. and Benoit, C., 1996], in which acoustical signal is generated by a Text to Speech engine and the phoneme information extracted from input text is used to define the articulatory movements.

To generate realistic facial animation is necessary to reproduce the contextual variability due to the reciprocal influence of articulatory movements for the production of following phonemes. This phenomenon, defined co-articulation [Farnetani and Recasens, 1999], is extremely complex and difficult to model. A variety of coarticulation strategies are possible and even different strategies may be needed for different languages [Bladon and Al-Bamerni, 1976]. A modified version of the Cohen-Massaro co-articulation model [Cosi and Perin, 2002] has been adopted for LUCIA [Fusaro et al., 2003] and a semi-automatic minimization technique, working on real cinematic data acquired by the ELITE opto-electronic system [Ferrigno and Pedotti, 1985], was used for training the dynamic characteristics of the model, in order to be more accurate in reproducing the true human lip movements.

Moreover, emotions are quite important in human interpersonal relations and individual development. Linguistic, paralinguistic and emotional transmission are inherently multimodal, and different types of information in the acoustic channel integrate with information from various other channels facilitating communicative processes. The transmission of emotions in speech communication is a topic that has recently received considerable attention, and automatic speech recognition (ASR) and multimodal or audio-visual (AV) speech synthesis are examples of fields, in which the processing of emotions can have a great impact and can improve the effectiveness and of human-machine interaction.

Viewing the face improves significantly the intelligibility of both natural and synthetic speech, especially under degraded acoustic conditions. Facial expressions signal emotions, add emphasis to the speech and

facilitate the interaction in a dialogue situation. From these considerations, it is evident that, in order to create more natural talking heads, it is essential that their capability comprises the emotional behavior.

In our TTS (text-to-speech) framework, AV speech synthesis, that is the automatic generation of voice and facial animation from arbitrary text, is based on parametric descriptions of both the acoustic and visual speech modalities. The visual speech synthesis uses 3D polygon models, that are parametrically articulated and deformed, while the acoustic speech synthesis uses an Italian version of the FESTIVAL diphone TTS synthesizer [Tesser, F. et al., 2001] now modified with emotive/expressive capabilities. The block diagram of our framework is depicted in Fig. 1a.

Various applications can be conceived by the use of animated characters, spanning from research on human communication and perception, via tools for the hearing impaired, to spoken and multimodal agent-based user interfaces. The recent introduction of WebGL, which is 3D graphics in web browsers, opens the possibility to bring all these applications via internet. A software porting of LUCIA facial animation framework is currently in development.

Figure 1. a) LUCIA's functional block diagram, b) position of reflecting markers and reference planes for the articulatory movement data collection (*on the left*), and the MPEG-4 standard facial reference points (*on the right*)

2. DATA ACQUISITION ENVIRONMENT

LUCIA is totally based on true real human data collected during the last decade by the use of ELITE [Cosi and Magno-Caldognetto, 1996 - Zmarich et al.,1997 - Magno-Caldognetto et al, 1998], a fully automatic movement analyzer for 3D kinematics data acquisition, which provides for 3D coordinate reconstruction, starting from 2D perspective projections, by means of a stereo-photogrammetric procedure which allows a free positioning of the TV cameras. The 3D data dynamic coordinates of passive markers (see Fig. 1b) are then used to create our lips articulatory model and to drive directly our talking face, copying human facial movements.

Two different configurations have been adopted for articulatory data collection: the first one, specifically designed for the analysis of labial movements, considers a simple scheme with only 8 reflecting markers (bigger markers in Fig.1b) while the second, adapted to the analysis of expressive and emotive speech, utilizes the full and complete set of 28 markers. All the movements of the 8 or 28 markers, depending on the adopted acquisition pattern, are recorded and collected, together with their velocity and acceleration, simultaneously with the co-produced speech which is usually segmented and analyzed by means of PRAAT [Boersma, 1996], that computes also intensity, du-ration, spectrograms, formants, pitch synchronous F0, and various voice quality parameters in the case of emotive and expressive speech [Magno Caldognetto et al., 2003 - Drioli et al., 2003].

In order to simplify and automates many of the operation needed for building-up the 3D avatar from the motion-captured data at ISTC-CNR we developed INTERFACE [Tisato et al, 2005], an integrated software designed and implemented in Matlab©

3. SYSTEM ARCHITECTURE

LUCIA is a graphic MPEG-4 standard compatible facial animation engine implementing a decoder compatible with the "Predictable Facial Animation Object Profile". LUCIA speaks with the Italian version of FESTIVAL TTS, as illustrated in Fig. 1a.

MPEG4 specifies a set of Face Animation Parameters (FAPs), each corresponding to a particular facial action deforming a face model in its neutral state. A particular facial action sequence is generated by deforming the face model, in its neutral state, according to the specified FAP values, indicating the magnitude of the corresponding action, for the corresponding time instant. Then the model is rendered onto the screen.

LUCIA is able to generate a 3D mesh polygonal model by directly importing its structure from a VRML file [Hartman and Wernecke, 1996] and to build its animation in real time.

Currently the model is divided in two sub sets of fundamental polygons: the skin on one hand and the inner articulators, such as the tongue and the teeth, or the facial elements such as the eyes and the hair, on the other. This subdivision is quite useful when animation is running, because only the reticule of polygons corresponding to the skin is directly driven by the pseudo-muscles and it constitutes a continuous and unitary element, while the other anatomical components move themselves independently and in a rigid way, following translations and rotations (for example the eyes rotate around their center). According to this strategy the polygons are distributed in such a way that the resulting visual effect is quite smooth with no rigid "jumps" over all the 3D model.

LUCIA emulates the functionalities of the mimic muscles, by the use of specific "displacement functions" and of their following action on the skin of the face. The activation of such functions is determined by specific parameters that encode small muscular actions acting on the face, and these actions can be modified in time in order to generate the wished animation. Such parameters, in MPEG-4, take the name of Facial Animation Parameters and their role is fundamental for achieving a natural movement. The muscular action is made explicit by means of the deformation of a polygonal reticule built around some particular key points called "Facial Definition Parameters" (FDP) that correspond to the junction on the skin of the mimic muscles.

Moving only the FDPs is not sufficient to smoothly move the whole 3D model, thus, each "feature point" is related to a particular "influence zone" constituted by an ellipses that represents a zone of the reticule where the movement of the vertexes is strictly connected. Finally, after having established the relationship for the whole set of FDPs and the whole set of vertexes, all the points of the 3D model can be simultaneously moved with a graded strength following a raised-cosine function rule associated to each FDP.

There are two current versions of LUCIA: an open source 3D facial animation framework written in C programming language (see http://sourceforge.net/projects/lucia/) and a new WebGL implementation. The C framework allows efficient rendering of a 3D face model in OpenGL-enabled systems (it has been tested on Windows and Linux using several architectures), has a modular design and provides several common facilities needed to create a real-time Facial Animation application. The very recent introduction of 3D graphics in the web browsers (which is known as WebGL) opens new possibilities for our 3D avatar. The powerful of this new technology is that you don’t need to download any additional software or driver to access the content of the 3D world you are interacting with. We are currently developing this new software version in order to easily integrate LUCIA in a website and use her as a virtual guide for the wikimemo.it project - The portal of Italian Language and Culture.

4. EMOTIONAL SYNTHESIS

Audio Visual emotional rendering was developed working on true real emotional audio and visual databases whose content was used to automatically train emotion specific intonation and voice quality models to be included in FESTIVAL Italian TTS system [Tesser et al., 2004 - Tesser et al, 2005 - Drioli et al., 2005.- Nicolao et al., 2006] and also to define specific emotional visual rendering to be implemented in LUCIA [Cosi et al., 2004 - Magno-Caldognetto et al., 2004 - Cavicchio et al, 2004].

An emotion specific XML editor explicitly designed for emotional tagged text was developed. The APML mark up language [Pelachaud et al., 2004] for behavior specification permits to specify how to markup the verbal part of a dialog move so as to add to it the "meanings" that the graphical and the speech generation components of an animated agent need to produce the required expressions.

So far, the language defines the components that may be useful to drive a face animation through the facial description language (FAP) and facial display functions. The extension of such language is intended to support voice specific controls. An extended version of the APML language has been included in the FESTIVAL speech synthesis environment, allowing the automatic generation of the extended phonation file

from an APML tagged text with emotive tags. This module implements a three-level hierarchy in which the affective high level attributes (e.g. <anger>, <joy>, <fear>) are described in terms of medium-level voice quality attributes defining the phonation type (e.g., <modal>, <soft>, <pressed>, <breathy>, <whispery>, <creaky>). These medium-level attributes are in turn described by a set of low-level acoustic attributes defining the perceptual correlates of the sound (e.g. <spectral tilt>, <shimmer>, <jitter>). The low-level acoustic attributes correspond to the acoustic controls that the extended MBROLA synthesizer can render through the sound processing procedure described above. This descriptive scheme has been implemented within FESTIVAL as a set of mappings between high-level and low-level descriptors. The implementation includes the use of envelope generators to produce time curves of each parameter.

In order to check and evaluate, by direct low-level manual/graphic instructions, various multi level emotional facial configurations we developed an EmotionPlayer, which was strongly inspired by the EmotionDisc of Zsofia Ruttkay [Ruttkay et al., 2003]. It is designed for a useful immediate feedback, as exemplified in Fig. 2.

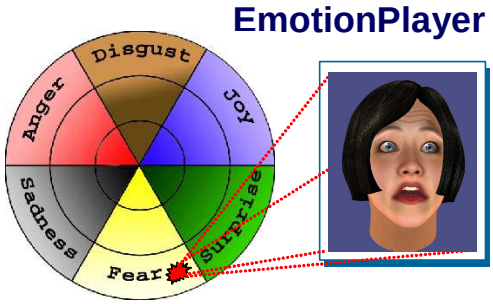

Figure 2. Emotion Player: clicking on three-level intensity(low, mid, high) emotional disc, an emotional configuration (i.e. high -fear) is activated.

5. CONCLUSION

LUCIA is an MPEG-4 standard FAPs driven OpenGL framework which provides several common facilities needed to create a real-time facial animation application. It has high quality 3D model and a fine coarticulation model, which is automatically trained by real data, used to animate the face.

The modified coarticulatory model is able to reproduce quite precisely the true cinematic movements of the articulatory parameters. The mean error between real and simulated trajectories for the whole set of parameters is, in fact, lower than 0.3 mm.

Labial movements implemented with the new modified model are quite natural and convincing especially in the production of bilabials and labiodentals and remain coherent and robust to speech rate variations.

The overall quality and user acceptability of LUCIA talking head has to be perceptually evaluated [Massaro, 1997 - Costantini et al., 2004] by a complete set of test experiments, and the new model has to be trained and validated in asymmetric contexts. Moreover, emotions and the behavior of other articulators, such as tongue for example, have to be analyzed and modeled for a better realistic implementation.

A new WebGL implementation of the avatar is currently in progress to exploit new possibilities that arise from the integration of LUCIA in the internet websites.

REFERENCES

Massaro, D.W. et al., 2000. Developing and Evaluating Conversational Agents. *Embodied Conversational Agents,* MIT Press, Cambridge, USA, pp. 287-318.

Bregler, C.et al., 1997. Video Rewrite: Driving Visual Speech with Audio. *Proceedings of SIGGRAPH '97*, pp. 353-360.

Terzopoulos, D. et al., 1995. Realistic Face Modeling for Animation. *Proceedings of SIGGRAPH '95*, pp. 55-62.

Vatikiotis-Bateson, E. et al.,1996. Physiology-Based Synthesis of Audiovisual Speech. *Proceedings of 4th Speech Production Seminar: Models and Data*, pp. 241-244.

Beskow, J., 1995. Rule-Based Visual Speech Synthesis. *Proceedings of Eurospeech '95*, Madrid, Spain, pp.299-302.

LeGoff, B. and Benoit, C., 1996. A text-to-audiovisual speech synthesizer for French. *Proceedings of ICSLP '96*, Philadelphia, USA, pp. 2163-2166.

Farnetani, E. and Recasens, D., 1999. Co-articulation Models in Recent Speech Production Theories. *Co-articulation in Speech Production*. Cambridge University Press, Cambridge, USA

Bladon, R.A. and Al-Bamerni, A., 1976. Co-articulation resistance in English. *Journal of Phonetics.* No.4, pp. 135-150

Cosi, P. and Perin, G., 2002. Labial Co-articulation Modeling for Realistic Facial Animation. *Proceedings of ICMI '02*, Pittsburgh, USA, pp. 505-510.

Fusaro, A. et al., 2003. LUCIA a New Italian Talking-Head Based on a Modified Cohen-Massaro's Labial Co-articulation Model. *Proceedings of Eurospeech 2003*, Geneva, Switzerland, vol. III, pp. 2269-2272.

Ferrigno, G. and Pedotti, A., 1985. ELITE: A Digital Dedicated Hardware System for Movement Analysis via Real-Time TV Signal Processing. *In IEEE Transactions on Biomedical Engineering*, BME-32, pp. 943-950.

Tesser, F. et al., 2001. Festival Speaks Italian!. *Proceedings of Eurospeech 2001*, Aalborg, Denmark, pp. 509-512.

Cosi, P., and Magno-Caldognetto, E., 1996. Lip and Jaw Movements for Vowels and Consonants: Spatio-Temporal Characteristics and Bimodal Recognition Applications. *In Speechreading by Humans and Machine: Models, Systems and Applications*.NATO ASI Series, Series F: Computer and Systems Sciences, vol. 150, pp. 291-313.

Zmarich, C. et al.,1997. Italian Consonantal Visemes: Relationships Between Spatial/temporal Articulatory Characteristics and Coproduced Acoustic Signal. *Proceedings of AVSP '97*, Rhodes, Greece, pp. 5-8.

Magno-Caldognetto, E. et al, 1998. Statistical Definition of Visual Information for Italian Vowels and Consonants. *Proceedings of AVSP '98*, Terrigal, Austria, pp. 135-140.

Boersma, P., 1996. PRAAT, a system for doing phonetics by computer. *In Glot International*, No. 5, pp. 341-345.

Magno Caldognetto, E. et al., 2003. Coproduction of Speech and Emotions: Visual and Acoustic Modifications of Some Phonetic Labial Tar-gets. *Proceedings of AVSP 2003*, St Jorioz, France, pp. 209-214.

Drioli, C. et al., 2003. Emotions and Voice Quality: Experiments with Sinusoidal Modeling. *Proceedings of Voqual 2003*, Geneva, Switzerland, pp. 127-132.

Tisato, G. et al, 2005. INTERFACE: a New Tool for Building Emotive/Expressive Talking Heads. *Proceedings of INTERSPEECH 2005*, Lisbon Portugal, pp. 781-784.

Hartman, J. and Wernecke, J., 1996. *The VRML Handbook*. Addison Wessley.

Tesser, F. et al., 2004. Prosodic Data-Driven Modelling of Narrative Style in FESTIVAL TTS. *Proceedings of 5th ISCA Speech Synthesis Workshop*, Pittsburgh, USA.

Tesser, F. et al, 2005. Emotional Festival-Mbrola TTS Synthesis. *Proceedings of INTERSPEECH 2005*, Lisbon, Portugal, pp. 505-508.

Drioli, C. et al., 2005. Control of Voice Quality for Emotional Speech Synthesis. *Proceedings of AISV 2004*, Padova, Italy, pp. 789-798.

Nicolao, M. et al., 2006. GMM modelling of voice quality for FESTIVAL-MBROLA emotive TTS synthesis. *Proceedings of INTERSPEECH 2006*, Pittsburgh, USA, pp. 1794-1797.

Cosi, P. et al., 2004. Data-Driven Tools for Designing Talking Heads Exploiting Emotional Attitudes. *Proceedings of Tutorial and Research Workshop "Affective Dialogue Systems"*, Kloster Irsee, Germany, pp. 101-112.

Magno-Caldognetto, E. et al., 2004. Visual and acoustic modifications of phonetic labial targets in emotive speech: Effects of the co-production of speech and emotions. *In Journal of Speech Communication*, vol. 44, pp. 173-185.

Cavicchio, F. et al, 2004. Modification of the Speech Articulatory Characteristics in the Emotive Speech. *Proceedings of Tutorial and Research Workshop "Affective Dialogue Systems"*, Kloster Irsee, Germany, pp. 233--239.

Pelachaud, C. et al., 2004. APML, a Mark-up Language for Believable Behavior Generation. *In Life-Like Characters*, Springer, pp. 65-85.

Ruttkay, Z. et al., 2003. Emotion Disc and Emotion Squares: tools to explore the facial expression space. *In Computer Graphics Forum*, No. 22, pp. 49-53

Massaro D.W., 1997: Perceiving Talking Faces. *Speech Perception to a Behavioral Principle*. MIT Press, Cambridge, USA

Costantini, E. et al., 2004. Evaluation of Synthetic Faces: Human Recognition of Emotional Facial Displays. *Proceedings of Tutorial and Research Workshop "Affective Dialogue Systems"*, Kloster Irsee, Germany, pp. 276-287.