EMOTIONAL TALKING HEAD: THE DEVELOPMENT OF "LUCIA"

Piero Cosi, Emanuela Caldognetto Magno, Graziano Tisato[1]

Istituto di Scienze e Tecnologie della Cognizione, CNR

ABSTRACT

The aim of this work is that of reviewing the ISTC activities focused on the development and implementation of LUCIA a new Italian emotional talking head based on a modified version of the Cohen-Massaro's labial coarticulation model.

LUCIA is an MPEG-4 standard facial animation system working on standard FAP visual parameters and speaking with the Italian version of FESTIVAL TTS. LUCIA was developed with the help of INTERFACE a Matlab© integrated software created to speed-up the procedure for building an emotive/expressive talking head from real human data. Various processing tools, working on dynamic articulatory data physically extracted by an optotracking 3D movement analyzer called ELITE, were implemented to build the animation engine and also to create the correct WAV and FAP files needed for the animation. By the use of INTERFACE, LUCIA, our animated MPEG-4 talking face, can copy a real human by reproducing the movements of passive markers positioned on his face and recorded by an opto-electronic device, or can be directly driven by an emotional XML tagged input text, thus realizing a true audio/visual emotive/expressive synthesis. LUCIA's voice is based on the ISTC Italian version of FESTIVAL - MBROLA packages, modified for expressive/emotive synthesis by means of an appropriate APML/VSML tagged language.

Index Terms— Emotions, Talking Head, TTS, 3D Facial Animation, LUCIA, FESTIVAL

1. INTRODUCTION

There are many ways to control a synthetic talking face. Among them, geometric parameterization [1-2], morphing between target speech shapes [3], muscle and pseudo-muscle models [4-5], appear the most attractive.

Recently, growing interest have encountered text to audiovisual systems [6-7], 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.

For generating 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 coarticulation [8], is extremely complex and difficult to model. A variety of coarticulation strategies are possible and even different strategies may be needed for different languages [9].

A modified version of the Cohen-Massaro coarticulation model [10] has been adopted for LUCIA [11] and a semi-automatic minimization technique, working on real cinematic data acquired by the ELITE opto-electronic system [12], was used for training the dynamic characteristics of the model, in order to be more accurate in reproducing the true human lip movements .

Morteover, 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 naturalness of man-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

[1] With the collaboration of many students working at ISTC during these last years, among them: Fabio Tesser, Carlo Drioli, Vincenzo Ferrari, Giulio Perin, Andrea Fusaro, Daniele Grigoletto, Mauro Nicolao, Giacomo Sommavilla, Enrico Marchetto …

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 [13] now modified with emotive/expressive capabilities.

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.

2. A/V DATA ACQUISITION ENVIRONMENT

LUCIA is totally based on true real human data collected during the last decade by the use of ELITE [14, 15, 16], a fully automatic movement analyzer for 3D kinematics data acquisition [12], which provides for 3D coordinate reconstruction (see Fig. 1), starting from 2D perspective projections, by means of a stereophotogrammetric procedure which allows a free positioning of the TV cameras.

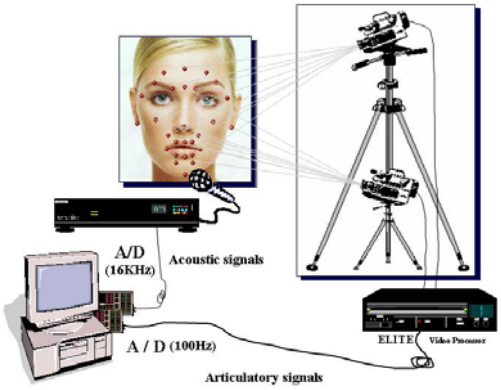

Figure 1. A/V acquisition environment.

The 3D data dynamic coordinates of passive markers such as those illustrated in Fig.2 are then used to create our lips articulatory model and to drive directly, copying human facial movements, our talking face.

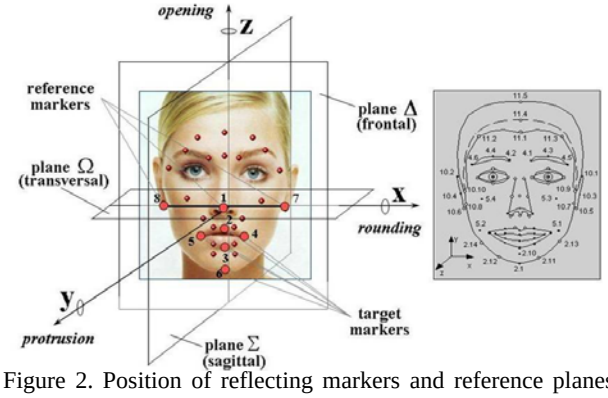

Figure 2. Position of reflecting markers and reference planes for the articulatory movement data collection (on the left), and the MPEG-4 standard facial reference points (on the right).

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 grey markers in Fig. 2) 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 [17], that computes also intensity, duration, spectrograms, formants, pitch synchronous F0, and various voice quality parameters in the case of emotive and expressive speech [18-19].

3. THE DEVELOPMENT ENVIRONMENT: INTERFACE

INTERFACE [20], whose block diagram is given in Figure 3, is an integrated software designed and implemented in Matlab© in order to simplify and automates many of the operation needed for building-up a talking head from motion-captured data. INTERFACE was mainly focused on articulatory data collected by ELITE, a fully automatic movement analyzer for 3D kinematics data acquisition [11], but could be easily adapted to other motion-captured data. ELITE provides for 3D coordinate reconstruction, starting from 2D perspective projections, by means of a stereophotogrammetric procedure which allows a free positioning of the TV cameras. The 3D data coordinates are then used to create our lips articulatory model and to drive directly, copying human facial movement, our talking face.

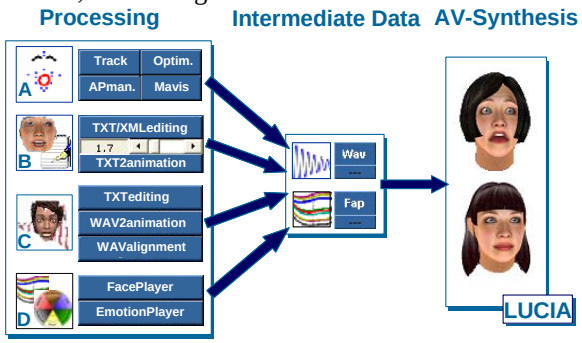

Figure 3. Block diagram of INTERFACE (see text for details).

INTERFACE was created mainly to develop LUCIA [11] our graphic MPEG-4 [21] compatible facial animation engine (FAE). In MPEG-4 FDPs (Facial Definition Parameters) define the shape of the model while FAPs (Facial Animation Parameters), define the facial actions [22]. In our case, the model uses a pseudo-muscular approach, in which muscle contractions are obtained through the deformation of the polygonal mesh around

feature points that correspond to skin muscle attachments. 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.

For a complete description of all the features and characteristics of INTERFACE see the official web site: http://www.pd.istc.cnr.it/INTERFACE. INTERFACE, handles four types of input data from which the corresponding MPEG-4 compliant FAP-stream could be created:

(A) Articulatory data, represented by the markers trajectories captured by ELITE; these data are processed by 4 programs:
- "**Track**", which defines the pattern utilized for acquisition and implements a new 3D trajectories reconstruction procedure;
- "**Optimize**", that trains the modified coarticulation model [6] utilized to move the lips of LUCIA, our current talking head under development;
- "**APmanager**", that allows the definition of the articulatory parameters in relation with marker positions, and that is also a DB manager for all the files used in the optimization stages;

(B) Symbolic high-level TXT/XML text data, processed by:
- "TXT/XMLediting", an emotional specific XML editor for emotion tagged text to be used in TTS and Facial Animation output;
- "TXT2animation", the main core animation tool that transforms the tagged input text into corresponding WAVand FAP files, where the first are synthesized by emotive/expressive FESTIVAL and the last, which are needed to animate MPEG-4 engines such as LUCIA, by the optimized animation model (designed by the use of Optimize);
- "TXTediting", a simple text editor for unemotional text to be used in TTS and Facial Animation output;

(C) WAV data, processed by:
- "WAV2animation", a simple tool that builds animations on the basis of input wav files after automatically segmenting them by an automatic ASR alignment system [8];
- "WAalignment", a simple segmentation editor to manipulate segmentation boundaries created by WAV2animation;

(D) manual graphic low-level data , created by:
- "FacePlayer", a direct low-level manual/graphic control of a single (or group of) FAP parameter; in other words, FacePlayer renders LUCIA's animation while acting on MPEG-4 FAP points for a useful immediate feedback;
- "EmotionPlayer", a direct low-level manual/graphic control of multi level emotional facial configurations for a useful immediate feedback.

4. LUCIA

LUCIA is the ISTC Italian talking head. LUCIA is based on the MPEG-4 standard [21] and speaks with the Italian version of FESTIVAL TTS [23], as illustrated in the block diagram shown in Figure 4.

LUCIA is a graphic MPEG-4 compatible facial animation engine implementing a decoder compatible with the "Predictable Facial Animation Object Profile" [21].

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.

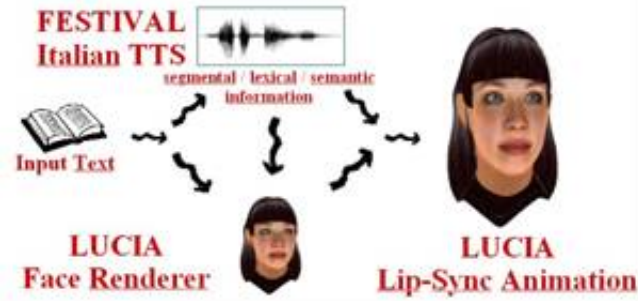

Figure 4: Lucia, a new Italian Talking head

LUCIA is able to generate a 3D mesh polygonal model by directly importing its structure from a VRML file [24] and to build its animation in real time.

At the current stage of development, as illustrated in Figure 5, LUCIA is a textured young female 3D face model built with 25423 polygons: 14116 belong to the skin, 4616 to the hair, 2688x2 to the eyes, 236 to the tongue and 1029 to the teeth respectively.

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.

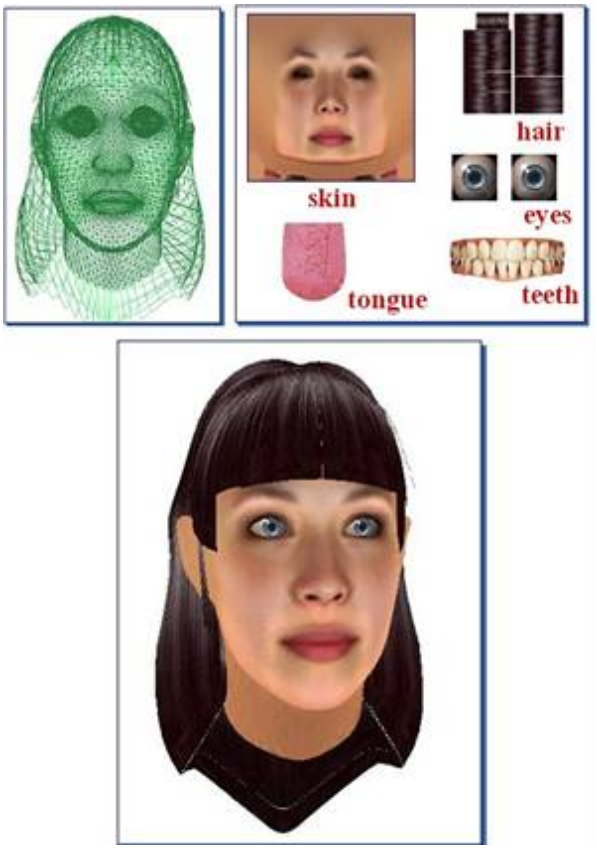

Figure 5: Lucia's wireframe and textures.

5. A/V 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, our Italian TTS system [25-28] and also to define specific emotional visual rendering to be implemented in LUCIA [29-31].

An emotion specific XML editor explicitly designed for emotional tagged text was developed. The APML mark up language [32] 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 (see Figure 6). 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 .pho 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>, etc.) are described in terms of medium-level voice quality attributes defining the phonation type (e.g., <modal>, <soft>, <pressed>, <breathy>, <whispery>, <creaky>, etc.). 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>, etc.). 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..

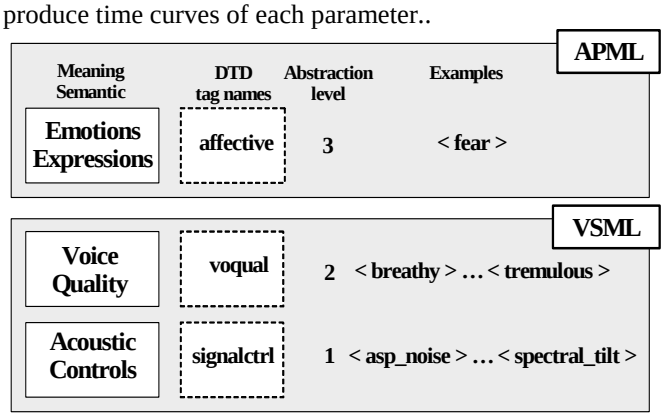

Figure 6: APML/VSML mark-up language extensions for emotive audio/visual synthesis.

Also an EmotionPlayer, which was strongly inspired by the EmotionDisc of Zsofia Ruttkay [33]), was developed in order to check end evaluate, by direct low-level manual/graphic instructions, various multi level emotional

facial configurations for a useful immediate feedback, as exemplified in Figure 7.

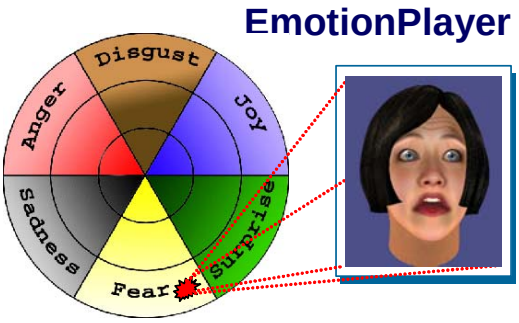

Figure 7: Emotion Player. Clicking on 3-level intensity (low, mid, high) emotional disc [33], an emotional configuration (i.e. high -fear) is activated.

6. CONCLUDING REMARKS

The graphic engine of LUCIA is similar to others MPEG based projects that were previously realized, but the novelty is the high quality of the 3D model, and the very 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 [34-35] by a complete set of test experiments, and the new model has to be trained and validated in asymmetric contexts (V1CV2) too. Moreover, emotions and the behavior of other articulators, such as tongue for example, have to be analyzed and modeled for a better realistic implementation.

Moreover, by the use of INTERFACE, the development of Facial Animation Engines and in general of expressive and emotive Talking Head could be made, and indeed it was for LUCIA, much more friendly. Evaluation tools will be included in the future such as, for example, perceptual tests for comparing human and talking head animations, thus giving us the possibility to get some insights about where and how the animation engine could be improved.

7. ACKNOWLEDGEMENTS

None of these studies, findings and applications would have been possible without the guide and sincere friendliness of TONI whose lost is still an open hurt.

Part of this work has been sponsored by PF-STAR (Preparing Future multiSensorial inTerAction Research, European Project IST- 2001-37599, http://pfstar.itc.it) and TICCA (Tecnologie cognitive per l'Interazione e la Cooperazione Con Agenti artificiali, joint "CNR - Provincia Autonoma Trentina" Project).

8. REFERENCES

[1] Massaro D.W., Cohen M.M., Beskow J., Cole R.A., "Developing and Evaluating Conversational Agents", in Cassell J., Sullivan J., Prevost S., Churchill E. (Editors), *Embodied Conversational Agents*, MIT Press, Cambridge, MA, 2000, pp. 287-318.

[2] Le Goff, B. *Synthèse à partir du texte de visages 3D parlant français*, PhD thesis, Grenoble, France, October 1997.

[3] Bregler C., Covell M., Slaney M., "Video Rewrite: Driving Visual Speech with Audio", in *Proceedings of SIGGRAPH '97*, 1997, pp. 353-360.

[4] Lee Y., Terzopoulos D., Waters K., "Realistic Face Modeling for Animation", in *Proceeding. of SIGGRAPH '95*, 1995, pp. 55–62.

[5] Vatikiotis-Bateson E., Munhall K.G., Hirayama M., Kasahara Y., Yehia H., "Physiology-Based Synthesis of Audiovisual Speech", in *Proceedings of 4th Speech Production Seminar: Models and Data*, 1996, pp. 241-244.

[6] Beskow J., "Rule-Based Visual Speech Synthesis," in *Proceedings of Eurospeech '95*, Madrid, 1995, pp.299–302.

[7] LeGoff B. and Benoit C., "A text-to-audiovisualspeech synthesizer for French", in *Proceedings of the ICSLP '96*, Philadelphia, USA, pp. 2163-2166.

[8] Farnetani E., Recasens D., "Coarticulation Models in Recent Speech Production Theories", in Hardcastle W.J. (Editors), *Coarticulation in Speech Production*, Cambridge University Press, Cambridge, 1999.

[9] Bladon, R.A., Al-Bamerni, A., "Coarticulation resistance in English \l\", *Journal of Phonetics*, 4, 1976, pp. 135-150.

[10] Cosi P., Perin G., "Labial Coarticulation Modeling for Realistic Facial Animation", in *Proceedings of ICMI '02*, Pittsburgh, PA, USA, 2002, pp. 505-510.

[11] Cosi P., Fusaro A., Tisato G., "LUCIA a New Italian Talking-Head Based on a Modified Cohen-Massaro's Labial Coarticulation Model", in *Proceedings of Eurospeech 2003*, Geneva, Switzerland, September 1–4, 2003, Vol. III, pp. 2269-2272.

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

[13] Cosi P., Tesser F., Gretter R., Avesani, C., "Festival Speaks Italian!", *in Proceedings of Eurospeech 2001*, Aalborg, Denmark, September 3-7, 509-512, 2001.

[14] Cosi P. and E. Magno Caldognetto E., "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*, D.G. Storke and M.E. Henneke eds., NATO ASI Series, Series F: Computer and Systems Sciences, Vol. 150, Springer-Verlag, 1996, pp. 291-313.

[15] Magno Caldognetto E., Zmarich C., Cosi P. and Ferrero F., "Italian Consonantal Visemes: Relationships Between Spatial/temporal Articulatory Characteristics and Coproduced Acoustic Signal", in *Proceedings of AVSP '97*, Tutorial & Research Workshop on Audio-Visual Speech Processing: Computational & Cognitive Science Approaches, Rhodes (Greece), 26-27 September 1997, pp. 5-8.

[16] Magno Caldognetto E., Zmarich C. and Cosi P., "Statistical Definition of Visual Information for Italian Vowels and Consonants", in D. Burnham, J. Robert-Ribes & E. Vatikiotis-Bateson (eds.), in *Proceedings of AVSP '98*, 4-6 December 1998, Terrigal (AUS), 5-7 Dec., 1998, pp. 135-140.

[17] Boersma P., "PRAAT, a system for doing phonetics by computer", *Glot International*, 5 (9/10), 341-345, 1996.

[18] Magno Caldognetto E., Cosi P., Drioli C., Tisato G., Cavicchio F., "Coproduction of Speech and Emotions: Visual and Acoustic Modifications of Some Phonetic Labial Tar-gets", in *Proceedings of AVSP 2003*, Audio Visual Speech Processing, ISCA Workshop, St Jorioz, France, September 4-7, 209-214, 2003.

[19] Drioli C., Tisato G., Cosi P., Tesser F., "Emotions and Voice Quality: Experiments with Sinusoidal Modeling", in *Proceedings of Voqual 2003*, Voice Quality: Functions, Analysis and Synthesis, ISCA Workshop, Geneva, Switzerland, August 27-29, 127-132, 2003.

[20] "INTERFACE: a New Tool for Building Emotive/Expressive Talking Heads", Tisato G., Cosi P., Drioli C., Tesser F., *in CD Proceedings INTERSPEECH 2005*, Lisbon, Portugal, 2005, pp. 781-784. (pdf)

[21] MPEG-4 standard. Home page: http://www.chiariglione.org/mpeg/index.htm

[22] Ekman P. and Friesen W., *Facial Action Coding System*, Consulting Psychologist Press Inc., Palo Alto (CA) (USA), 1978.

[23] Cosi P., Tesser F., Gretter R., Avesani C., "Festival Speaks Italian!", in *Proceedings of Eurospeech 2001*, Aalborg, Denmark, September 3-7 2001, pp. 509-512.

[24] Hartman J., Wernecke J., *The VRML Handbook*, Addison Wessley, 1996.

[25] Tesser F., Cosi P., Drioli C., Tisato G., "Prosodic Data-Driven Modelling of Narrative Style in FESTIVAL TTS", in *CD Proceedings of 5th ISCA Speech Synthesis Workshop*, 14th-16th June 2004, Carnegie Mellon University, Pittsburgh USA.

[26] Tesser F., Cosi P., Drioli C., Tisato G., "Emotional Festival-Mbrola TTS Synthesis", , in *CD Proceedings INTERSPEECH 2005*, Lisbon, Portugal, 2005, pp. 505-508.

[27] "Control of Voice Quality for Emotional Speech Synthesis", Drioli C., Tesser F., Tisato G., Cosi P., in *CD* Proceedings of AISV 2004, 1st Conference of Associazione Italiana di Scienze della Voce, Padova, Italy, December 2-4, 2004, EDK Editore s.r.l., Padova, 2005, pp. 789-798.

[28] Nicolao M., Drioli C., Cosi P., "GMM modelling of voice quality for FESTIVAL/MBROLA emotive TTS synthesis", in *Proceedings of INTERSPEECH 2006*, Pittsburgh, Pennsylvania, USA, 17-21 September, 2006, pp. 1794-1797.

[29] Cosi P., Fusaro A., Grigoletto D., Tisato G., "Data-Driven Tools for Designing Talking Heads Exploiting Emotional Attitudes", in *Proceedings of Tutorial and Research Workshop "Affective Dialogue Systems"*, June 14 - 16, 2004, Kloster Irsee, Germany, pp. 101-112.

[30] Magno Caldognetto E., Cosi P., Drioli C., Tisato G., Cavicchio F., "Visual and acoustic modifications of phonetic labial targets in emotive speech: Effects of the co-production of speech and emotions", Speech Communication, Vol. 44, October 2004, pp. 173-185.

[31] Magno Caldognetto E., Cosi P., Cavicchio F., "Modification of the Speech Articulatory Characteristics in the Emotive Speech", in *Proceedings of Tutorial and Research Workshop "Affective Dialogue Systems"*, June 14 - 16, 2004, Kloster Irsee, Germany, pp. 233-239. (pdf)

[32] De Carolis, B., Pelachaud, C., Poggi I., and Steedman M., "APML, a Mark-up Language for Believable Behavior Generation", in Prendinger H., Ishizuka M. (eds.), *Life-Like Characters*, Springer, pp. 65-85, 2004.

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

[34] Massaro D.W., *Perceiving Talking Faces: from Speech Perception to a Behavioral Principle*, Cambridge, MA, MIT Press, 1997.

[35] Costantini E., Pianesi F., Cosi P., "Evaluation of Synthetic Faces: Human Recognition of Emotional Facial Displays", in Proceedings of Tutorial and Research Workshop "Affective Dialogue Systems", June 14 - 16, 2004, Kloster Irsee, Germany, pp. 276-287.