LUCIA: An Open Source 3D Expressive Avatar for Multimodal h.m.i.

G. Riccardo Leone, Giulio Paci, and Piero Cosi

Institute of Cognitive Sciences and Technologies – National Research Council
Via Martiri della libertà 2, 35137 Padova, Italy
{piero.cosi,riccardo.leone}@pd.istc.cnr.it

Abstract. LUCIA is an MPEG-4 facial animation system developed at ISTC-CNR[1]. 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 built from real human data physically extracted by ELITE optic-tracking movement analyzer. LUCIA can copy a real human being 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 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, TTS, facial animation, mpeg4, 3D avatar, virtual agent, affective computing, 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.

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.

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

---

[1] With the collaboration of many students and researchers working at ISTC during these last years, among them: G.Tisato, F.Tesser, C.Drioli, V.Ferrari, G.Perin, A.Fusaro, D.Grigoletto, M.Nicolao, G.Sommavilla, E.Marchetto.

A modified version of the Cohen-Massaro co-articulation model [10] has been adopted for LUCIA [11] and a semi-automatic minimization technique, working on real cinematic data acquired by the ELITE optic-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.

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 the 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 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.
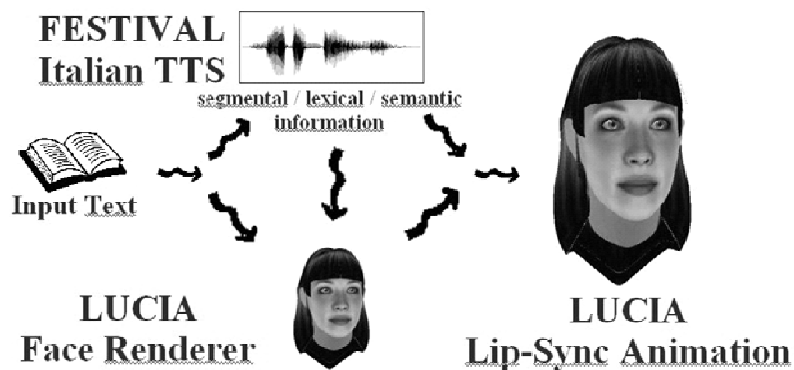


Fig. 1. LUCIA's functional block diagram

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 [13] now modified with emotive/expressive capabilities. The block diagram of our framework is depicted in Fig. 1.

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 [14], which is 3D graphics in web browsers, opens the possibility to bring all these applications via internet. This software version of LUCIA is currently in the early stage of development.

2   Data Acquisition Environment

LUCIA is totally based on true real human data collected during the last decade by the use of ELITE [15, 16, 17], a fully automatic movement analyzer for 3D cinematics data acquisition [12], which provides 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. 2) are then used to create our lips articulatory model and to drive directly our talking face, copying human facial movements.
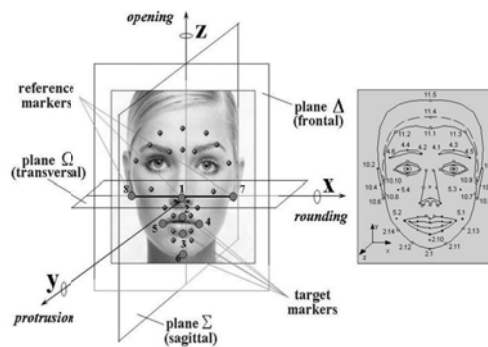
Fig. 2. Position of reflecting markers and reference planes for the articulatory movement data collection (*on the left*), and the MPEG-4 standard facial reference points (*on the right*)

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 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 [18], that computes also intensity, du-ration, spectrograms, formants, pitch synchronous F0, and various voice quality parameters in the case of emotive and expressive speech [19, 20].

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 [21], an integrated software designed and implemented in Matlab©.

3   Architecture and Implementations

LUCIA is a MPEG-4 standard facial animation engine implementing a decoder compatible with the "Predictable Facial Animation Object Profile" [22]. LUCIA speaks with the Italian version of FESTIVAL TTS [13]; we have already seen the overall system illustrated in Fig. 1. The homepage of the project is [23].

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, 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 [24] and to build its animation in real time.

At the current stage of development, 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.
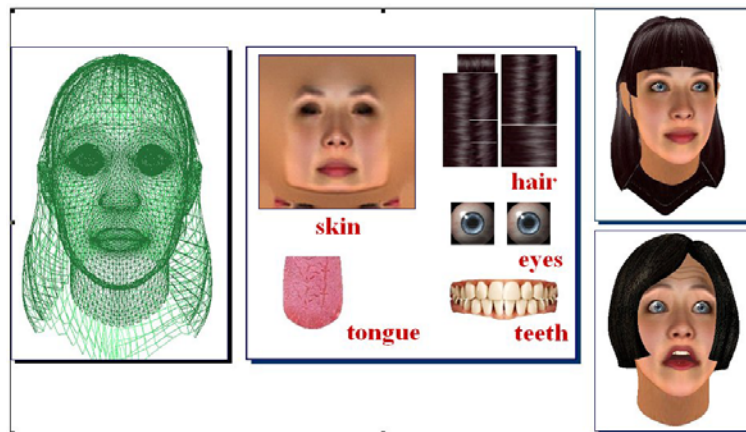
Fig. 3. Lucia's wireframe, textures and renderings

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 constitutes a continuous and unitary element, while the other anatomical components move themselves independently, 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; 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 (FDPs) 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 [25] and a new WebGL implementation [26]. 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). It has a modular design: each module provides one of the several common facilities needed to create a real-time Facial Animation application. It includes an interface to play audio files, a robust and extendable FAP parser, sample-based audio/video synchronization and an object oriented interface to the components of a face model. Finally, the framework includes a ready to use female face model and some sample applications to play simple animation and to test movements' behavior. The framework's core is built on top of OpenGL and does not rely on any specific context provider (it has been tested using GLUT[27], FreeGLUT[28] and GtkGLExt[29]). In order to grant portability, the modules' interfaces are designed so that their implementation details are hidden to the application and it is possible to provide multiple implementation of the same model (e.g., three implementations of the audio module are included using respectively OpenAL[30], Gstreamer[31] and Microsoft MCI[32]). The very recent introduction of 3D graphics in the web browsers (which is known as WebGL [14]) 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; there are many promising functionality for web applications: a virtual guide (which we are exploiting in the Wikimemo.it project - The portal of Italian Language and Culture); a storyteller for e-book reading; a digital tutor for hearing impaired; a personal assistant for smart-phone and mobile devices. The early results can be observed in [26].

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, our Italian
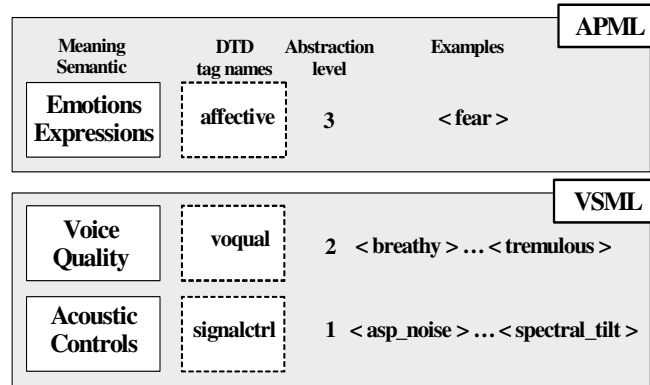
Fig. 4. APML/VSML mark-up language extensions for emotive audio/visual synthesis

TTS system [33, 34, 35, 36] and also to define specific emotional visual rendering to be implemented in LUCIA [37, 38, 39].

An emotion specific XML editor explicitly designed for emotional tagged texts was developed. The APML mark up language [40] 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 (Fig. 4). 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 "EmotionPlayer", which was strongly inspired by the EmotionDisc of Zsofia Ruttkay [41]. It is designed for a useful immediate feedback, as exemplified in Fig. 5.
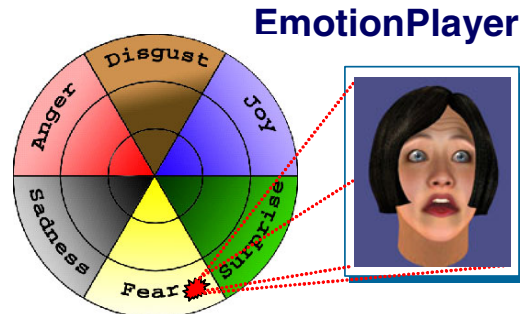
EmotionPlayer

Fig. 5. Emotion Player: clicking on three-level intensity (low, mid, high) emotional disc, an emotional configuration (i.e. high -fear) is activated

5   Conclusions

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 co-articulatory model, which is automatically trained by real data, used to animate the face.

The modified co-articulatory 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 [42, 43] 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.

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

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. (eds.) Embodied Conversational Agents, pp. 287–318. MIT Press, Cambridge (2000)
2. Le Goff, B.: Synthèse à partir du texte de visages 3D parlant français. PhD thesis, Grenoble, France (1997)
3. Bregler, C., Covell, M., Slaney, M.: Video Rewrite: Driving Visual Speech with Audio. In: SIGGRAPH 1997, pp. 353–360 (1997)
4. Lee, Y., Terzopoulos, D., Waters, K.: Realistic Face Modeling for Animation. In: SIGGRAPH 1995, pp. 55–62 (1995)
5. Vatikiotis-Bateson, E., Munhall, K.G., Hirayama, M., Kasahara, Y., Yehia, H.: Physiology-Based Synthesis of Audiovisual Speech. In: 4th Speech Production Seminar: Models and Data, pp. 241–244 (1996)
6. Beskow, J.: Rule-Based Visual Speech Synthesis. In: Eurospeech 1995, Madrid, Spain, pp.299–302 (1995)
7. LeGoff, B., Benoit, C.: A text-to-audiovisualspeech synthesizer for French. In: ICSLP 1996, Philadelphia, U.S.A, pp. 2163–2166 (1996)
8. Farnetani, E., Recasens, D.: Coarticulation Models in Recent Speech Production Theories. In: Hardcastle, W.J. (ed.) Coarticulation in Speech Production. Cambridge University Press, Cambridge (1999)
9. Bladon, R.A., Al-Bamerni, A.: Coarticulation resistance in English. J. Phonetics. 4, 135–150 (1976)
10. Cosi, P., Perin, G.: Labial Coarticulation Modeling for Realistic Facial Animation. In: ICMI 2002, Pittsburgh, U.S.A, pp. 505–510 (2002)
11. Cosi, P., Fusaro, A., Tisato, G.: LUCIA a New Italian Talking-Head Based on a Modified Cohen-Massaro's Labial Coarticulation Model. In: Eurospeech 2003, Geneva, Switzerland, vol. III, pp. 2269–2272 (2003)
12. Ferrigno, G., Pedotti, A.: ELITE: A Digital Dedicated Hardware System for Movement Analysis via Real-Time TV Signal Processing. IEEE Transactions on Biomedical Engineering, BME-32, 943–950 (1985)
13. Cosi, P., Tesser, F., Gretter, R., Avesani, C.: Festival Speaks Italian! In: Eurospeech 2001, Aalborg, Denmark, pp. 509–512 (2001)
14. WebGL- OpenGL for the web, `http://www.khronos.org/webgl/`
15. Cosi, P., Magno Caldognetto, E.: Lip and Jaw Movements for Vowels and Consonants: Spatio-Temporal Characteristics and Bimodal Recognition Applications. In: Storke, D.G., Henneke, M.E. (eds.) Speechreading by Humans and Machine: Models, Systems and Applications. NATO ASI Series, Series F: Computer and Systems Sciences, vol. 150, pp. 291–313. Springer, Heidelberg (1996)
16. Magno Caldognetto, E., Zmarich, C., Cosi, P., Ferrero, F.: Italian Consonantal Visemes: Relationships Between Spatial/temporal Articulatory Characteristics and Coproduced Acoustic Signal. In: AVSP 1997, Tutorial & Research Workshop on Audio-Visual Speech Processing: Computational & Cognitive Science Approaches, Rhodes, Greece, pp. 5–8 (1997)
17. Magno Caldognetto, E., Zmarich, C., Cosi, P.: Statistical Definition of Visual Information for Italian Vowels and Consonants. In: Burnham, D., Robert-Ribes, J., Vatikiotis-Bateson, E. (eds.) Proceedings of AVSP 1998, Terrigal, Austria, pp. 135–140 (1998)

18. Boersma, P.: PRAAT, a system for doing phonetics by computer. Glot International 5(9/10), 341–345 (1996)
19. Magno Caldognetto, E., Cosi, P., Drioli, C., Tisato, G., Cavicchio, F.: Coproduction of Speech and Emotions: Visual and Acoustic Modifications of Some Phonetic Labial Targets. In: AVSP 2003, ISCA Workshop, St Jorioz, France, pp. 209–214 (2003)
20. 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, pp. 127–132 (2003)
21. Tisato, G., Cosi, P., Drioli, C., Tesser, F.: INTERFACE: a New Tool for Building Emotive/Expressive Talking Heads. In: INTERSPEECH 2005, Lisbon, Portugal, pp. 781–784 (2005)
22. MPEG-4 standard,
    `http://mpeg.chiariglione.org/standards/mpeg-4/mpeg-4.htm`
23. LUCIA - homepage, `http://www2.pd.istc.cnr.it/LUCIA/`
24. Hartman, J., Wernecke, J.: The VRML Handbook. Addison Wessley (1996)
25. LUCIA Open source project, `http://sourceforge.net/projects/lucia/`
26. LUCIA WebGL version, `http://www2.pd.istc.cnr.it/LUCIA/webgl/`
27. The OpenGL Utility Toolkit,
    `http://www.opengl.org/resources/libraries/glut/`
28. The Free OpenGL Utility Toolkit, `http://freeglut.sourceforge.net`
29. GTK+ OpenGL Extension, `http://projects.gnome.org/gtkglext/`
30. OpenAL:a cross platform 3D audio API,
    `http://connect.creativelabs.com/openal/`
31. Gstreamer:Open source multimedia framework,
    `http://gstreamer.freedesktop.org/`
32. Windows MCI,
    `http://en.wikipedia.org/wiki/Media_Control_Interface`
33. Tesser, F., Cosi, P., Drioli, C., Tisato, G.: Prosodic Data-Driven Modelling of Narrative Style in FESTIVAL TTS. In: 5th ISCA Speech Synthesis Workshop, Pittsburgh, U.S.A (2004)
34. Tesser, F., Cosi, P., Drioli, C., Tisato, G.: Emotional Festival-Mbrola TTS Synthesis. In: INTERSPEECH 2005, Lisbon, Portugal, pp. 505–508 (2005)
35. Drioli, C., Tesser, F., Tisato, G., Cosi, P.: Control of Voice Quality for Emotional Speech Synthesis. In: 1st Conference of Associazione Italiana di Scienze della Voce, AISV 2004, EDK Editore s.r.l., Padova, Italy, pp. 789–798 (2005)
36. Nicolao, M., Drioli, C., Cosi, P.: GMM modelling of voice quality for FESTIVAL-MBROLA emotive TTS synthesis. In: INTERSPEECH 2006, Pittsburgh, U.S.A, pp. 1794–1797 (2006)
37. Cosi, P., Fusaro, A., Grigoletto, D., Tisato, G.: Data-Driven Tools for Designing Talking Heads Exploiting Emotional Attitudes. In: André, E., Dybkjær, L., Minker, W., Heisterkamp, P. (eds.) ADS 2004. LNCS (LNAI), vol. 3068, pp. 101–112. Springer, Heidelberg (2004)
38. 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. J. Speech Communication 44, 173–185 (2004)
39. Magno Caldognetto, E., Cosi, P., Cavicchio, F.: Modifications of Speech Articulatory Characteristics in the Emotive Speech. In: André, E., Dybkjær, L., Minker, W., Heisterkamp, P. (eds.) ADS 2004. LNCS (LNAI), vol. 3068, pp. 233–239. Springer, Heidelberg (2004)

40. De Carolis, B., Pelachaud, C., Poggi, I., Steedman, M.: APML, a Mark-up Language for Believable Behavior Generation. In: Prendinger, H., Ishizuka, M. (eds.) Life-Like Characters, pp. 65–85. Springer, Heidelberg (2004)
41. Ruttkay, Z., Noot, H., Hagen, P.: Emotion Disc and Emotion Squares: tools to explore the facial expression space. Computer Graphics Forum 22(1), 49–53 (2003)
42. Massaro, D.W.: Perceiving Talking Faces: from Speech Perception to a Behavioral Principle. MIT Press, Cambridge (1997)
43. Costantini, E., Pianesi, F., Cosi, P.: Evaluation of Synthetic Faces: Human Recognition of Emotional Facial Displays. In: André, E., Dybkjær, L., Minker, W., Heisterkamp, P. (eds.) ADS 2004. LNCS (LNAI), vol. 3068, pp. 276–287. Springer, Heidelberg (2004)