Data-Driven Tools for Designing Talking Heads Exploiting Emotional Attitudes

Piero Cosi, Andrea Fusaro, Daniele Grigoletto, Graziano Tisato

Istituto di Scienze e Tecnologie della Cognizione - C.N.R.
Sezione di Padova "Fonetica e Dialettologia"
Via G. Anghinoni, 10 - 35121 Padova ITALY
{cosi, fusaro, grigoletto, tisato}@csrf.pd.cnr.it
http://www.csrf.pd.cnr.it

Abstract. Audio/visual speech, in the form of labial movement and facial expression data, was utilized in order to semi-automatically build a new Italian expressive and emotive talking head capable of believable and emotional behavior. The methodology, the procedures and the specific software tools utilized for this scope will be described together with some implementation examples.

1 Introduction[1]

It is quite evidently documented by specific workshops and conferences (AVSP, LREC), European (FP6, Multimodal/Mulsensorial Communication, R&D) and International (COCOSDA, ISLE, LDC, MITRE) framework activities, and by various questionnaires (see ISLE and NIMM, ELRA, COCOSDA, LDC, TalkBank, Dagstuhl Seminar [1]) that data-driven procedures for building more natural and expressive talking heads are becoming popular and successful.

The knowledge that both acoustic and visual signal simultaneously convey linguistic, extra linguistic and paralinguistic information is rather spread in the speech communication community, and it constitutes the basis for this work. The data-driven procedure utilized to build a new Italian talking head, described in this work, has been, in fact, directly driven by audio/visual data, in the form of labial movement and facial expression data, that were physically extracted by an automatic optotracking movement analyzer for 3D kinematics data acquisition called ELITE [2].

[1] Part of this work has been sponsored by COMMEDIA (COMunicazione Multimodale di Emozioni e Discorso in Italiano con Agente animato virtuale, CNR Project C00AA71), 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).

1.1 Audio/Visual Acquisition Environment

ELITE is a fully automatic movement analyzer for 3D kinematics data acquisition, that provides for 3D coordinate reconstruction, starting from the 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 calculate and evaluate the parameters described hereinafter. 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 on Figure 1a) while the second, adapted to the analysis of expressive and emotive speech, utilizes the full and complete set of 28 markers.

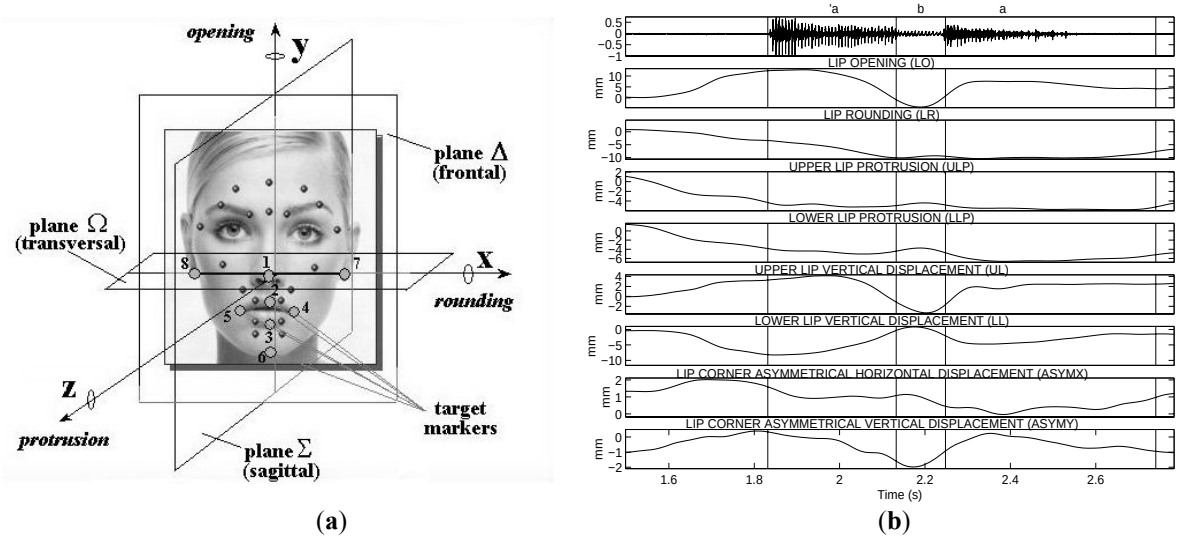

Fig. 1. Position of reflecting markers and reference planes for the articulatory movement data collection **(a)**; speech signal and time evolution of some labial kinematic parameters (LO, LR, ULP, LLP, UL, LL, ASYMX and ASYMY, see text) for the sequence /'aba/ **(b)**.

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 [3], that computes also intensity, duration, spectrograms, formants, pitch synchronous F0, and various voice quality parameters in the case of emotive and expressive speech [4-5]. As for the analysis of the labial movements, the most common parameters selected to quantify the labial configuration modifications, as illustrated in Figure 1b for some of them, are introduced in the following Table:

• **Lip Opening** (**LO**), calculated as the distance between markers placed on the central points of the upper and lower lip vermillion borders [d(m2,m3)]; this parameter correlates with the HIGH-LOW phonetic dimension.
• **Lip Rounding** (**LR**), corresponding to the distance between the left and right corners of the lips [d(m4,m5)], which correlates with the ROUNDED-UNROUNDED phonetic dimension: negative values correspond to the lip spreading.
• **Anterior/posterior movements** (Protrusion) **of Upper Lip and Lower Lip** (ULP and **LLP**), calculated as the distance between the marker placed on the central points of either the upper and lower lip and the frontal plane Δ containing the line crossing the markers placed on the lobes of the ears and perpendicular to Ω plane [d(m2,Δ), d(m3,Δ)]. These parameters correlate with the feature PROTRUDED-RETRACTED: negative values quantify the lip retraction.
• **Upper** and **Lower Lip vertical displacements** (**UL**, **LL**), calculated as a distance between the markers placed on the central point of either upper and lower lip and the transversal plane Ω passing through the tip of the nose and the markers on the ear lobes [d(m2,Ω), d(m3,Ω)]. Hence, positive values correspond to a reduction of the displacement of the markers from the Ω plane. As told before, these parameters are normalized in relation to the lip resting position.

Table 1. Meaning of some of the most common chosen articulatory parameters.

2 Data-Driven Methodology and Tools

As explained in [6-8], several Audio/Visual corpora, were used to train our MPEG-4 [9] standard talking head called LUCIA [10] speaking with an Italian version of FESTIVAL TTS [11].

2.1 Model estimation

The parameter estimation procedure for LUCIA's model is based on a least squared *phoneme-oriented* error minimization scheme with a strong convergence property, between real articulatory data *Y(n)* and modeled curves *F(n)* for the whole set of *R* stimuli belonging to the same phoneme set :

$$e = \sum_{r=1}^{R} \left(\sum_{n=1}^{N} \left(Y_r(n) - F_r(n) \right)^2 \right)$$

where *F(n)* is generated by a modified version of the Cohen-Massaro co-articulation model [13] as introduced in [6-7]. Even if the number of parameters to be optimized is rather high, the size of the data corpus is large enough to allow a meaningful estimation, but, due to the presence of several local minima, the optimization process has to be manually controlled in order to assist the algorithm convergence. The mean total error between real and simulated trajectories for the whole set of parameters is lower than 0.3 mm in the case of bilabial and labiodental consonants in the /a/ and /i/ contexts [14, p. 63].

2.2 MPEG4 Animation

In MPEG-4 [9], *FDPs* (*Facial Definition Parameters*) define the shape of the model while *FAPs* (*Facial Animation Parameters*), define the facial actions. Given the shape of the model, the animation is obtained by specifying the *FAP-stream* that is for each frame the values of FAPs (see Figure 2). In a *FAP-stream,* each frame has two lines of parameters. In the first line the activation of a particular marker is indicated (0, 1) while in the second, the target values, in terms of differences from the previous ones, are stored.

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

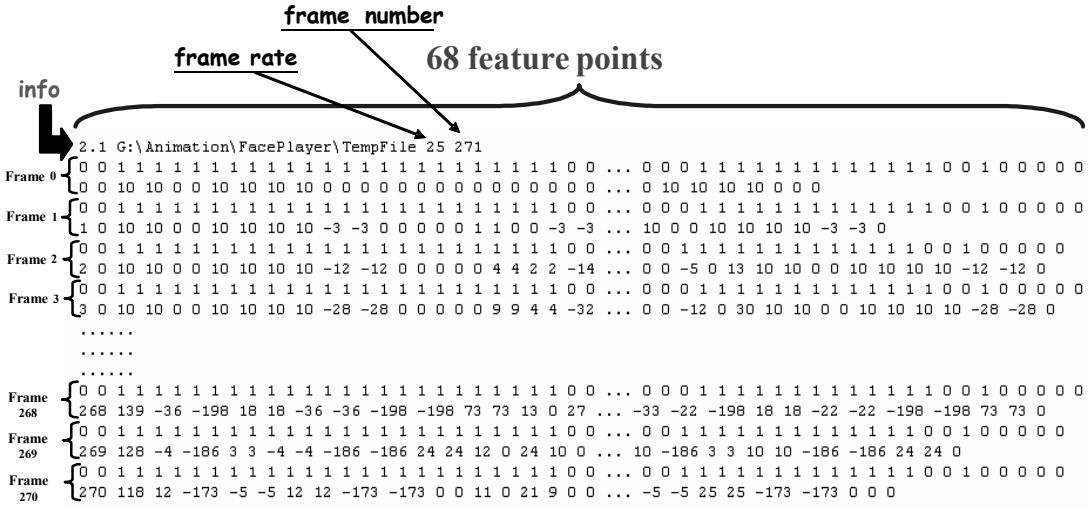

Fig. 2. The FAP stream.

Each feature point follows MPEG4 specifications where a FAP corresponds to a minimal facial action. When a FAP is activated (i.e. when its intensity is not null) the feature point on which the FAP acts is moved in the direction signalled by the FAP itself (up, down, left, right, etc).

Using the pseudo-muscular approach, the facial model's points within the region of this particular feature point get deformed. A facial expression is characterised not only by the muscular contraction that gives rise to it, but also by an intensity and a duration. The intensity factor is rendered by specifying an intensity for every FAP. The temporal factor is modelled by three parameters: onset, apex and offset [15].

The *FAP-stream* needed to animate a FAE (*Facial Animation Engine*) could be completely synthesized by using a specific animation model, such as the co-articulation one used in LUCIA, or it could be reconstructed on the basis of real data captured by an optotracking hardware, such as ELITE.

2.3 Tools: "INTERFACE"

In order to speed-up the procedure for building-up our talking head an integrated software called **INTERFACE,** whose block diagram is illustrated in Figure 3, was designed and implemented in Matlab©. INTERFACE simplifies and automates many of the operation needed for that purpose.

The whole processing block is designed in order to prepare the correct wav and FAP files needed for the animation engines, both in the sense of building up the engines and of truly creating the current wav and FAP file needed for the final animation. The final animation, in fact, can be completely synthesized starting from an input emotional tagged text, by the use of our animation engine [13], or it can be reproduced by using the data, relative to the specific movements of the markers positioned on human subjects, extracted by ELITE.

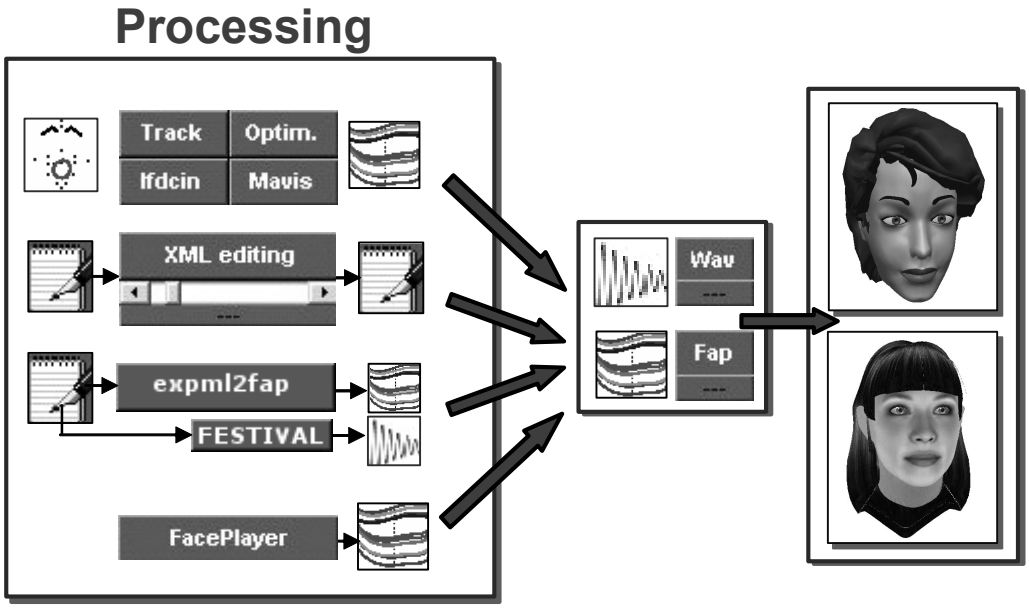

Fig. 3. INTERFACE block diagram (see text for details)

INTERFACE, handles three types of input data from which the corresponding MPEG4 compliant FAP-stream could be created:

- **low-level 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 the 3D trajectories reconstruction procedure;

- "**OPTIMIZE**" that trains the modified co-articulation model [13] utilized to move the lips of **GRETA** [6] and **LUCIA** [10], our two current talking heads under development;
- "**IFDCIN**", 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;
- "**MAVIS**" (*Multiple Articulator VISualizer,* written by Mark Tiede of ATR Research Laboratories [16]) that allows different visualizations of articulatory signals;

- **symbolic high-level XML text data,** processed by:
  - "**XML-EDITING**", an emotional specific XML editor for emotion tagged text to be used in TTS and Facial Animation output;
  - **"EXPML2FAP",** the main core animation tool that transforms the tagged input text into corresponding WAVand FAP files, where the first are synthesized by FESTIVAL and the last, which are needed to animate the MPEG4 engines GRETA or LUCIA [11], by the optimized animation model (designed by the use of OPTIMIZE);

- **single low-level FAPs**, created by:
  - "**XML-EDITING**", (see above);

  and edited by
  - "**FACEPLAYER**", a direct low-level manual control of a single (or group of) Fap; in other words, FACEPLAYER renders what happen, in GRETA and LUCIA, while acting on MPEG4 FAP points for a useful immediate feedback.

The *TrackLab* software originally supplied by BTS© [17] for ELITE is not reliable in reconstructing 3D trajectories when there are a lot of very quickly varying markers close to each other, as it usually happens in the articulatory study of facial expressions. The TRACK MatLab© software was, in fact, developed with the aim of avoiding marker tracking errors that force a long manual post-processing stage and also a compulsory stage of markers identification in the initial frame for each used camera. TRACK is quite effective in terms of trajectories reconstruction and processing speed, obtaining a very high score in marker identification and reconstruction by means of a reliable adaptive processing. Moreover only a single manual intervention for creating the reference tracking model (*pattern of markers*) is needed for all the files acquired in the same working session. TRACK, in fact, tries to guess the possible target pattern of markers, as illustrated in Figure 4, and the user must only accept a proposed association or modify a wrong one if needed, then it runs automatically on all files acquired in the same session.

Moreover, we let the user the possibility to independently configure the markers and also a standard FAP-MPEG. The actual configuration of the FAP is described in an initialization file and can be easily changed. The markers assignment to the MPEG standard points is realized with the context menu as illustrated in Figure 5.

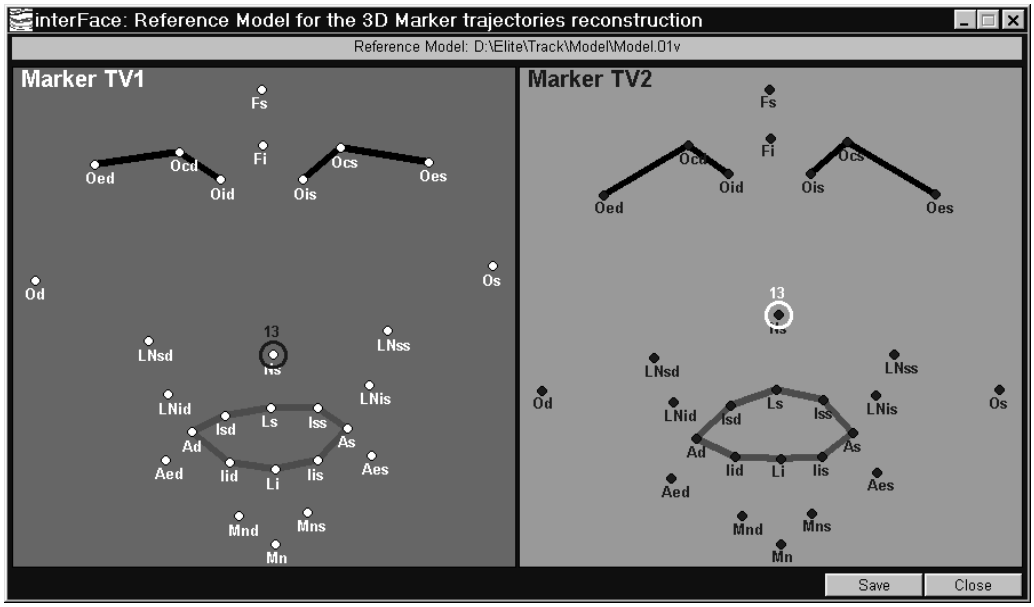

Fig. 4. Definition of the reference model. TRACK's marker positions and names are associated with those corresponding to the real case.

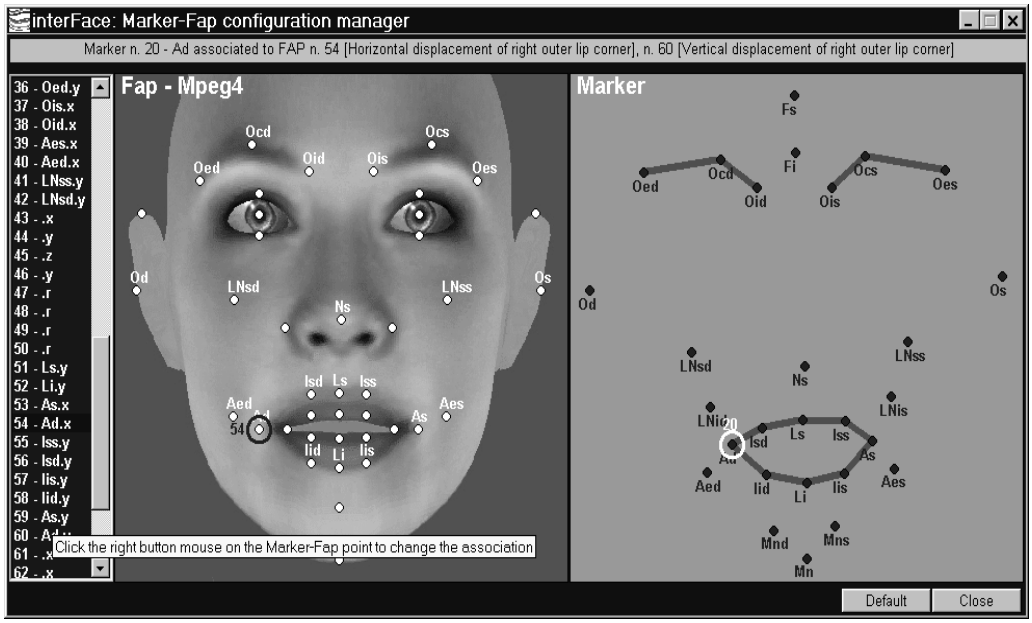

Fig. 5. Marker MPEG-FAP association with the TRACK's reference model. The MPEG reference points (on the left) are associated with the TRACK's marker positions (on the right).

In other words, as illustrated in the examples shown in Figure 6, for LUCIA, TRACK allows 3D real data driven animation of a talking face, converting the ELITE trajectories into standard MPEG4 data and eventually it allows, if necessary, an easy editing of bad trajectories. Different MPEG4 FAEs could obviously be animated with the same FAP-stream allowing for an interesting comparison among their different renderings.

Fig. 6. Examples of a single-frame LUCIA's emotive expressions. These were obtained by acquiring real movements with ELITE, by automatically tracking and reconstructing them with "TRACK" and by reproducing them with LUCIA.

3 Visual Emotions

At the present time, emotional visual configurations are designed and refined, by means of visual inspection of real data, with a software called **EMOTIONALPLAYER** (EP) (see Figure 7), designed and implemented in Matlab© on the basis of FACIALPLAYER, introduced above in 2.3, and greatly inspired by the Emotion Disc software [18]. In the future, a strategy similar to that introduced in 2.1 will be adopted. EMOTIONAL PLAYER manages single facial movements of a synthetic face in a standard MPEG-4 framework in order to create emotional and expressive visual renderings in GRETA and LUCIA.

As already underlined above in 2.2, in MPEG-4 animations, FDPs define the shape of the model while FAPs define the facial actions. The intensity and the duration of an emotive expression are driven by an intensity factor that is rendered by specifying an intensity for every FAP, and by a temporal factor which is modelled by onset, apex and offset parameters, as explained in [15].

The onset and offset represent, respectively, the time the expression takes to appear and to disappear; the apex corresponds to the duration for which the facial expression is at its peak intensity value. These parameters are fundamental to convey the proper meaning of the facial expressions. In our system, every facial expression is characterised by a set of FAPs. Every set of FAPs allows for example the creation of the 6 facial expressions corresponding to the 6 basic primary emotions of Ekman's set (Table 2), chosen here for a sake of simplicity, and for every expression only 3 levels of intensity (low, medium, high) have been simulated.

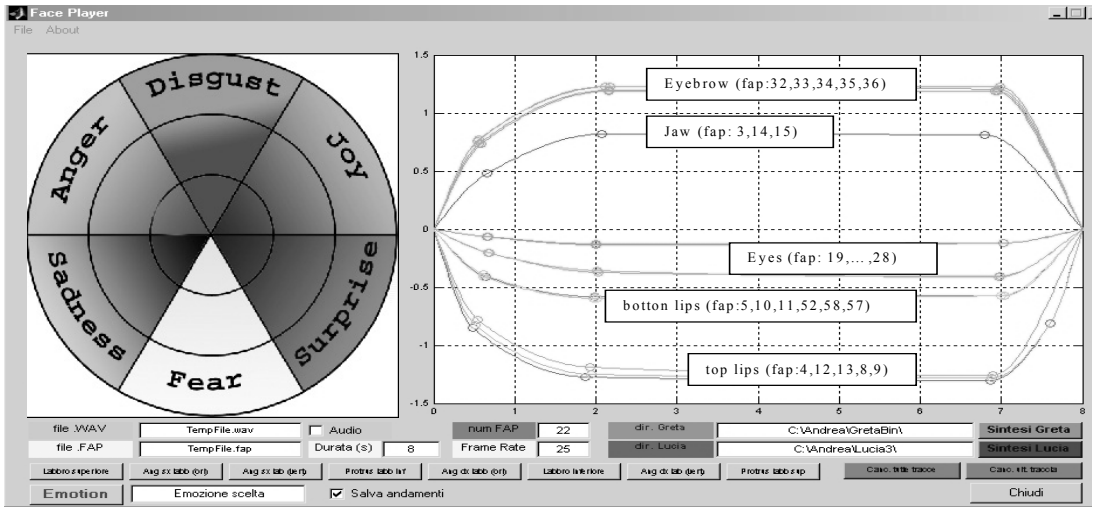

Fig. 7. EMOTIONALPLAYER.

Expression	Description
Anger	The inner eyebrows are pulled downward and together. The eyes are wide open. The lips are pressed against each other or opened to expose the teeth
Fear	The eyebrows are raised and pulled together. The inner eyebrows are bent upward. The eyes are tense and alert.
Disgust	The eyebrows and eyelids are relaxed. The upper lip is raised and curled, often asymmetrically.
Happiness	The eyebrows are relaxed. The mouth is open and the mouth corners pulled back toward the ears.
Sadness	The inner eyebrows are bent upward. The eyes are slightly closed. The mouth is relaxed.
Surprise	The eyebrows are raised. The upper eyelids are wide open, the lower relaxed. The jaw is opened

Table. 2. The 6 basic primary emotions of Ekman's set with corresponding facial expressions**.**

In our system we distinguish "emotion basis" *EB(t)* from "emotion display" *ED(t)*. They are both functions of the time *t*. An *EB(t)* involves a specific zone of the face such as the eyebrow, mouth, jaw, eyelid and so on. *EB(t)* includes also facial movements such as nodding, shaking, turning the head and movement of the eyes. Each *EB(t)* is defined as a set of MPEG-4 compliant FAP parameters:

$$EB(t) = \{ fap3 = v_1(t); .....; fap68 = v_k(t)\}$$

where $v_1(t),. . . ,v_k(t)$ specify the FAPs function intensity value created by the user. An *EB(t)* can also be defined as a combination of *EB'(t)* by using the '+' operator in this way:

$$EB'(t) = EB_1'(t) + EB_2'(t)$$

The emotion display is finally obtained by a linear scaling:

$$ED'(t) = EB(t)*c = \{ fap3 = v_1(t)*c; \ldots\ldots\ldots\ldots; fap68 = v_k(t)*c)\}$$

where EB is a "facial basis" and 'c' a constant. The operator '*' multiplies each of the FAPS constituting the EB by the constant 'c'. The onset, offset and apex (i.e. the duration of the expression) of emotion is determined by the weighed sum of the functions $v_k(t)$ (k = 3,...,68) created by mouse actions. In Figure 8, two simple emotional examples for fear and happiness are illustrated.

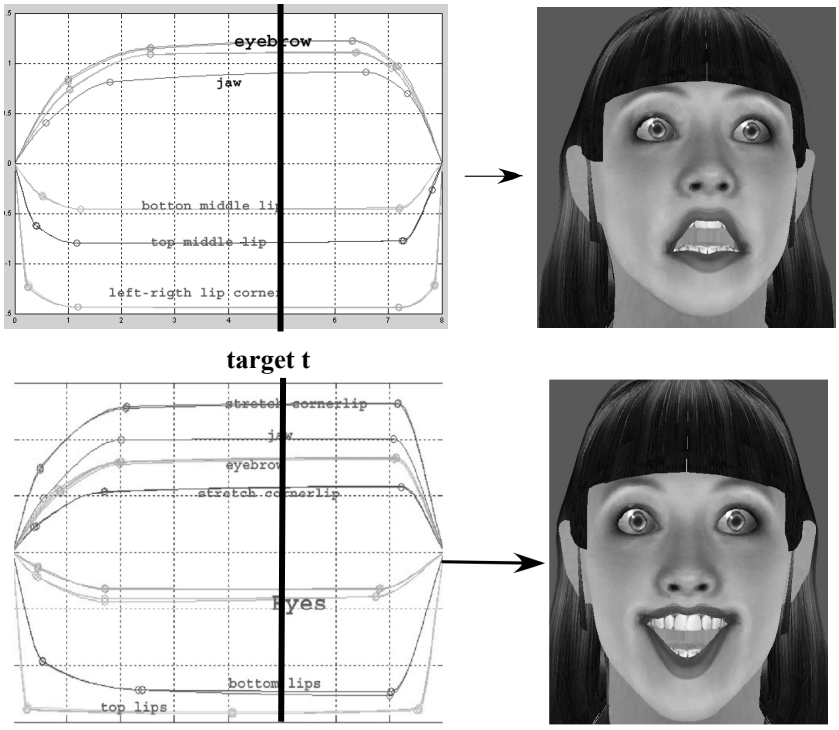

Fig. 8. Fear (top) and happiness (bottom) emotional examples.

4 Concluding Remarks

An integrated software environment designed and developed for the acquisition, creation, management, access, and use of audio/visual (AV) articulatory data, captured by an automatic optotracking movement analyzer, has been introduced and described in its general characteristics. These methods, tools, and procedures can surely accelerate the development of Facial Animation Engines and in general of expressive and emotive Talking Agents.

5 Future Trends

Evaluation should be strongly carried out in the future and evaluation tools will be included in these tools. Perceptual tests, for example, for comparing both the original videos/signals and the talking head can surely give us some insights about where and how the animation engine could be improved.

Results from a preliminar experiment for the evaluation of the adequacy of facial displays in the expression of some basic emotional states, based on a recognition task, are presented in a paper in this volume, where also the potentials of the used evaluation methodology are discussed [19].

References

1. Working Group at the Dagstuhl Seminar on Multimodality, 2001, questionnaire on Multimodality http://www.dfki.de/~wahlster/Dagstuhl_Multi_Modality
2. Ferrigno G., Pedotti A., "ELITE: A Digital Dedicated Hardware System for Movement Analysis via Real-Time TV Signal Processing", IEEE Trans. on Biomedical Engineering, BME-32, 1985, 943-950.
3. Boersma P., "PRAAT, a system for doing phonetics by computer", Glot International, 5 (9/10), 1996, 341-345.
4. 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", Proc. AVSP 2003, Audio Visual Speech Processing, ISCA Workshop, St Jorioz, France, September 4-7, 2003, 209-214 .
5. Drioli C., Tisato G., Cosi P., Tesser F., "Emotions and Voice Quality: Experiments with Sinusoidal Modeling", Proceedings of Voqual 2003, Voice Quality: Functions, Analysis and Synthesis, ISCA Workshop, Geneva, Switzerland, August 27-29, 2003, 127-132.
6. Pelachaud C., Magno Caldognetto E., Zmarich C., Cosi P., "Modelling an Italian Talking Head", Proc. AVSP 2001, Aalborg, Denmark, September 7-9, 2001, 72-77.
7. Cosi P., Magno Caldognetto E., Perin G., Zmarich C., "Labial Coarticulation Modeling for Realistic Facial Animation", Proc. ICMI 2002, 4th IEEE International Conference on Multimodal Interfaces 2002, October 14-16, 2002 Pittsburgh, PA, USA., pp. 505-510.
8. Cosi P., Magno Caldognetto E., Tisato G., Zmarich C., "Biometric Data Collection For Bimodal Applications", Proceedings of COST 275 Workshop, The Advent of Biometric on the Internet, November 7-8, 2002, Rome, pp. 127-130.
9. MPEG-4 standard. Home page: http://www.chiariglione.org/mpeg/index.htm.
10. Cosi P., Fusaro A., Tisato G., "LUCIA a New Italian Talking-Head Based on a Modified Cohen-Massaro's Labial Coarticulation Model", Proc. Eurospeech 2003, Geneva, Switzerland, September 1-4, 2003, 127-132.
11. Cosi P., Tesser F., Gretter R., Avesani, C., "Festival Speaks Italian!", Proc. Eurospeech 2001, Aalborg, Denmark, September 3-7, 2001, 509-512.
12. FACEGEN web page: http://www.facegen.com/index.htm.
13. Cohen M., Massaro D., "Modeling Coarticulation in Synthetic Visual Speech", in Magnenat-Thalmann N., Thalmann D. (Editors), Models and Techniques in Computer Animation, Springer Verlag, Tokyo, 1993, pp. 139-156.
14. Perin G., "Facce parlanti: sviluppo di un modello coarticolatorio labiale per un sistema di sintesi bimodale", MThesis, Univ. of Padova, Italy, 2000-1.

15. Ekman P. and Friesen W., Facial Action Coding System, Consulting Psychologist Press Inc., Palo Alto (CA) (USA), 1978.
16. Tiede, M.K., Vatikiotis-Bateson, E., Hoole, P. and Yehia, H, "Magnetometer data acquisition and analysis software for speech production research", ATR Technical Report TRH 1999, 1999, ATR Human Information Processing Labs, Japan.
17. BTS home page: http://www.bts.it/index.php
18. Ruttkay Zs., Noot H., ten Hagen P., "Emotion Disc and Emotion Squares: tools to explore the facial expression space", Computer Graphics Forum, 22(1) 2003, 49-53.
19. Costantini E., Pianesi F., Cosi P., "Evaluation of Synthetic Faces: Human Recognition of Emotional Facial Displays", (in this volume)