INTERFACE: a New Tool for Building Emotive/Expressive Talking Heads

Graziano Tisato, Piero Cosi, Carlo Drioli, Fabio Tesser

Istituto di Scienze e Tecnologie della Cognizione
Sezione di Padova "Fonetica e Dialettologia"
Consiglio Nazionale delle Ricerche
35121 Padova, ITALY
[tisato, cosi, drioli, tesser]@pd.istc.cnr.it

Abstract

In order to speed-up the procedure for building an emotive/expressive talking head such as LUCIA, an integrated software called INTERFACE was designed and implemented in Matlab©. INTERFACE simplifies and automates many of the operations needed for that purpose. A set of processing tools, focusing mainly on dynamic articulatory data physically extracted by an automatic optotracking 3D movement analyzer, was implemented in order to build up the animation engine, that is based on the Cohen-Massaro coarticulation model, and also to create the correct WAV and FAP files needed for the animation. LUCIA, our animated MPEG-4 talking face, in fact, can *copy* a real human by reproducing the movements of some markers positioned on his face and recorded by an optoelectronic 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 an Italian version of FESTIVAL - MBROLA packages, modified for expressive/emotive synthesis by means of an appropriate APML/VSML tagged language.

1. Introduction

Among scientists there is widespread agreement that emotions are important in human relations and individual development, that the linguistic and emotional transmission is inherently multimodal, and that different types of information in the acoustic channel integrate with information from various other channels. The transmission of emotions in speech communication is a topic that has recently received considerable attention. 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 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 emotional behaviour.
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 [1] 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.
The aim of this work was that of implementing INTERFACE a flexible architecture that allows us to easily develop and test a new animated face speaking in Italian.

2. INTERFACE

INTERFACE, whose block diagram is given in Figure 1, is an integrated software designed and implemented in Matlab© in order to simplify and automate many of the operations needed for building-up a talking head. INTERFACE is mainly focused on articulatory data collected by ELITE, a fully automatic movement analyzer for 3D kinematics data acquisition [2]. 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 movements, our talking face.
INTERFACE was created mainly to develop LUCIA [3], our graphic MPEG-4 [4] 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 [5]. 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, a full detailed PDF manual is being prepared and it is available at the official LUCIA web site:
http://www.pd.istc.cnr.it/LUCIA/Docs/InterFace-AISV2004.pdf

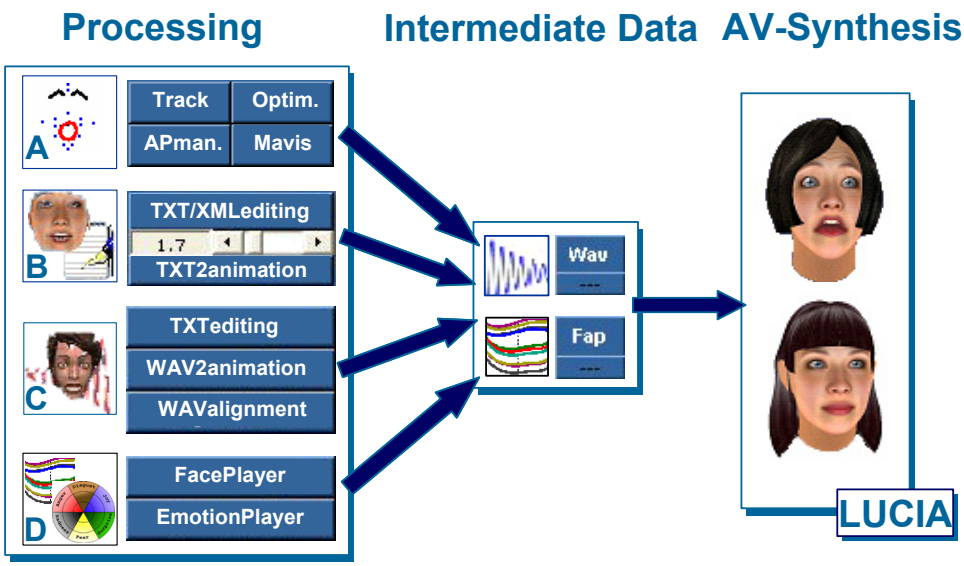

Figure 1: Block diagram of INTERFACE

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 infrared passive marker 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***", which trains the modified coarticulation model [6] utilized to move the lips of a MPEG-4 compliant talking face;
  - "***APmanager***", which allows the definition of the articulatory parameters in relation with marker positions, and that is also a database manager for all the files used in the optimization stages;
  - "***Mavis***" (Multiple Articulator VISualizer, written by Mark Tiede of ATR Research Laboratories [7]), which allows different visualizations of articulatory signals;
- **(B) Symbolic high-level TXT/XML text data**, processed by:
  - "***TXT/XMLediting***", a specific XML editor for emotive/expressive 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 WAV and FAP files. The audio file is synthesized by a FESTIVAL module, which realizes the emotive/expressive vocal modifications. The FAP-stream file, needed to animate MPEG-4 engines such as LUCIA, is obtained by an animation model, designed by the use of *Optimize*;
  - "***TXTediting***", a simple editor for text without any kind of tags, to be used in TTS and Facial Animation output;
- **(C) WAV data**, processed by:
  - "***WAV2animation***", a tool that builds animations on the basis of input WAV files after automatically segmenting them by an automatic ASR alignment system [8];
  - "***WAValignment***", 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 useful immediate feedback;
  - "***EmotionPlayer***", a direct low-level manual/graphic control of multi level emotional facial configurations for a useful immediate feedback.

2.1. "Track"

MatLab© *Track* was developed with the aim of avoiding marker tracking errors that force a long manual post-processing stage and also a compulsory stage of marker 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 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 give the user the possibility to independently configure the markers and also the FAP-MPEG correspondence. The actual configuration of the FAPs is described in an initialization file and can be easily changed. The markers assignment to MPEG standard points is realized with a context menu as illustrated in Figure 2. By *Track*, the articulatory movements can also be separated from the head roto-translation, thus allowing to realize a correct data driven articulatory synthesis.

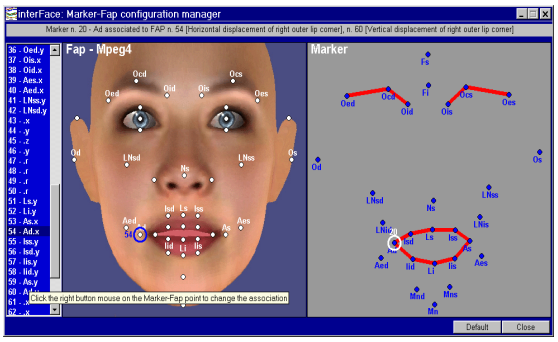

Figure 2: Marker MPEG-FAP association with the TRACK's reference model. The MPEG reference points (on the left) are associated with the TRACK's marker positions (on the right).

In other words , as illustrated in the examples shown in Figure 3, for LUCIA, *Track* allows 3D real data driven animation of a talking face, converting the ELITE trajectories into standard MPEG-4 data and eventually it allows, if necessary, an easy editing of *bad* trajectories. Different MPEG-4 Facial Animation Engines (FAEs) could obviously be animated with the same FAP-stream allowing for an interesting comparison among their different renderings.

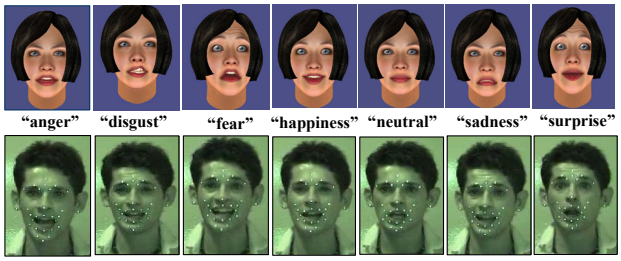

Figure 3: Examples of single-frame LUCIA's emotive expressions. These were obtained by acquiring real human movements with ELITE, by automatically tracking and reconstructing them with "Track", and by reproducing them with LUCIA.

2.2. "Optimize"

The *Optimize* module implements the parameter estimation procedure for LUCIA's lip articulation model. This procedure 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 coarticulation model [6] as introduced in [9-10]. 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 [11, p. 63]. At the end of the optimization stage, the lip movements of our MPEG-4 LUCIA can be obtained simply starting from a WAV file and its corresponding phoneme segmentation information.

2.3. "TXT/XMLediting"

This is an emotion specific XML editor explicitly designed for emotional tagged text. The APML mark up language [12] 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 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 .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.

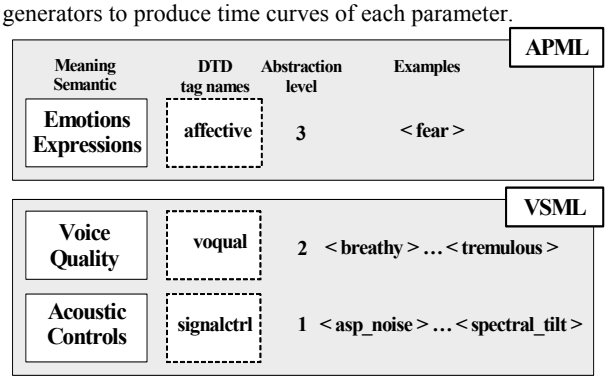

Figure 4: APML/VSML mark-up language extensions for emotive audio/visual synthesis.

2.4. "TXT2animation"

This represents the main animation module. *TXT2animation* transforms the emotional tagged input text into corresponding WAV and FAP files, where the first are synthesized by the Italian emotive version of FESTIVAL, and the last by the optimized coarticulation model, as for the lip movements, and

by specific facial action sequences obtained for each emotion by knowledge-based rules. For example, *anger* can be activated using knowledge-based rules acting on action units AU2 + AU4 + AU5 + AU10 + AU20 + AU24, where Action Units correspond to various facial action (i.e. AU1: "inner brow raiser", AU2: "outer brow raiser", etc.) [5]. MPEG-4 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. In other words, lips are animated by the use of the optimized data driven articulation model, while the full face is animated following knowledge-based rules.

2.5. "WAV2animation" and "WAVsegmentation"

WAV2animation is essentially similar to the previous *TXT2animation* module, but in this case an audio/visual animation is obtained starting from a WAV file instead that from a text file. An automatic segmentation algorithm based on a very effective Italian ASR system [8] extracts the phoneme boundaries. These data could be also verified and edited by the use of the *WAVsegmentation* module, and finally processed by the final visual only animation module of *TXT2animation*. At the present time, the animation is neutral because the data do not correspond to a tagged emotional text, but in future this option will be made available.

2.6. "FacePlayer" and "EmotionPlayer"

The first module *FacePlayer* lets the user verify immediately through the use of a direct low-level manual/graphic control of a single (or group of) FAP (acting on MPEG4 FAP points) how LUCIA renders the corresponding animation for a useful immediate feedback. *EmotionPlayer*, which was strongly inspired by the EmotionDisc of Zsofia Ruttkay [13]), is instead a direct low-level manual/graphic control of multi level emotional facial configurations for a useful immediate feedback, as exemplified in Figure 5.

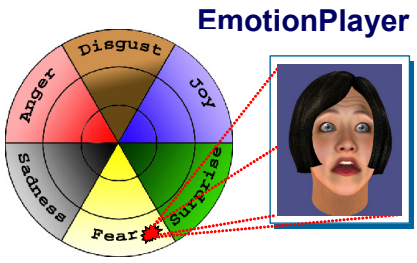

Figure 5: Emotion Player. Clicking on 3-level intensity (low, mid, high) emotional disc [13], an emotional configuration (i.e. high -fear) is activated.

3. Conclusions

With the use of INTERFACE, the development of Facial Animation Engines and in general of expressive and emotive Talking Agents 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.

4. Acknowledgements

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).

5. References

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

[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, pp. 943-950, 1985.

[3] 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, pp. 127-132, 2003.

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

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

[6] 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, pp. 139-156, 1993.

[7] 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*, ATR Human Information Processing Labs, Japan, 1999.

[8] Cosi P. and Hosom J.P., "High Performance 'General Purpose' Phonetic Recognition for Italian", Proc. of ICSLP 2000, Beijing, Cina, Vol. II, pp. 527-530, 2000.

[9] Pelachaud C., Magno Caldognetto E., Zmarich C., Cosi P., "Modelling an Italian Talking Head", *Proc. AVSP 2001*, Aalborg, Denmark, September 7-9, 2001, pp. 72-77.

[10] Cosi P., Magno Caldognetto E., Perin G., Zmarich C., "Labial Coarticulation Modeling for Realistic Facial Animation", *Proc.* 4th IEEE International Conference on Multimodal Interfaces *ICMI 2002*, Pittsburgh, PA, USA, pp. 505-510, 2000.

[11] Perin G., *Facce parlanti: sviluppo di un modello coarticolatorio labiale per un sistema di sintesi bimodale*, MThesis, Univ. of Padova, Italy, 2000-1.

[12] 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.

[13] 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.