Describing "*INTERFACE*" a Matlab© Tool for Building Talking Heads

Piero Cosi, Graziano Tisato

Istituto di Scienze e Tecnologie della Cognizione - Sede di Padova "Fonetica e Dialettologia"
Consiglio Nazionale delle Ricerche
piero.cosi@pd.istc.cnr.it, graziano.tisato@pd.istc.cnr.it

Abstract

INTERFACE is a Matlab©.tool for simplifying and automating many of the operations needed for building a talking head. INTERFACE consists of set of processing tools, focusing on dynamic articulatory data physically extracted by an automatic optotracking 3D movement analyzer. The final aim of INTERFACE is that of building up the animation engine of LUCIA our emotive/expressive Italian talking head. LUCIA can be directly driven by an emotional XML tagged input text, thus realizing a true audio visual expressive synthesis. LUCIA's voice is based on an Italian version of FESTIVAL-MBROLA packages, modified for expressive synthesis by means of an appropriate APML/VSML tagged language. Moreover, by using INTERFACE, it is possible to copy a real human talking by recreating the correct WAV and FAP files needed for the animation by reproducing the movements of some markers positioned on his face and recorded by an optoelectronic device. In this work the latest improvements of INTERFACE will be described and few examples of their application to real cases will be illustrated.

Index Terms: Talking Head, Audio/Visual synthesis, Articulation, Emotions.

1. Introduction

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 [2], whose initial screenshot 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.

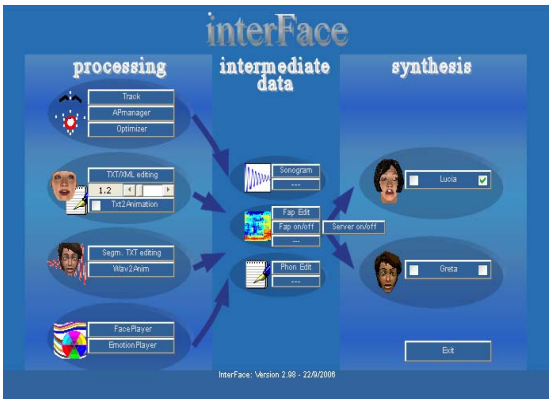

Figure 1: *INTERFACE starting screenshot.*

INTERFACE is mainly focused on articulatory data collected by ELITE, a fully automatic movement analyzer for 3D kinematics data acquisition [3], but it could be also used with similar data captured by analogous hardware instruments. 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 [4], our graphic MPEG-4 [5] 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 [6]. 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.

As illustrated in Figure1, INTERFACE is divided in three blocks: data "*processing*", editing and control of "*intermediate data*", and audio/visual "*synthesis*". The

"*options*" push-button in the top menu allows the user to configure the different modules and applications and to save initialization parameters in a specific initialization file.

The "*processing*" zone is the fundamental part of INTERFACE, and it contains different modules developed for dealing with data, subdivided according to their type, that is: bimodal data (articulatory data) , XML and textual data, audio and low level (FAP) data.

The functionalities of the "*intermediate data*" zone regard the visualization of the speech waveform, of its relative sonogram, and also the visualization and editing of the phonetic segmentation and of course of the FAP articulatory animation data. An important innovation, introduced in this last version of INTERFACE, is the mechanism of synchronization of real and synthetic animation video thus allowing the user to easily compare them.

As far as "*syntehsis*", the talking faces can be activated on the same computer or they can be started in a client/server modality thus transmitting and receiving the FAP animation data in a local net or within the web via a specific IP address.

The audio synthesis is generated by FESTIVAL [1] coupled with the classical MBROLA engine [7] and also with a new developed SMS (Spectral Modeling Synthesis) engine [8] for Italian.

In summary, 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 [9] 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;
  - "***Sonogram***" a new INTERFACE visualization module which is able to visualize, synchronously with video and articulatory signals, the speech waveform, its corresponding sonogram and pitch (see Figure 2). Moreover with this new feature it is also possible to edit and save articulatory parameters in order manually adjust the final animation when needed. This new feature substitutes the previous "***Mavis***" (Multiple Articulator VISualizer, written by Mark Tiede of ATR Research Laboratories [10]) module.
- **(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 [11];
  - "***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.

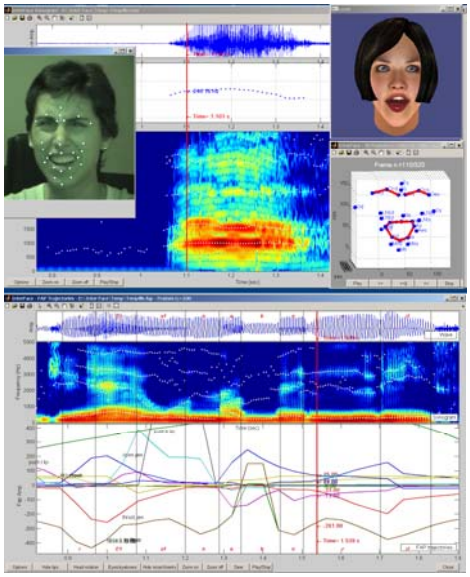

Figure 2: "***Sonogram***" module for INTREFACE.

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 3. By Track, the

articulatory movements can also be separated from the head roto-translation, thus allowing to realize a correct data driven articulatory synthesis.

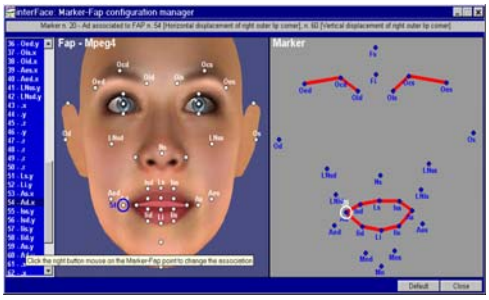

Figure 3: *Marker MPEG-FAP association with the TRACK's reference model. The MPEG reference points (on the left) are associated with the TRACK's marker positions (on the right).*

In other words , as illustrated in the examples shown in Figure 4, for LUCIA, Track allows a true 3D "*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.

Figure 4: *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.*

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.

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 [9] as introduced in [12-13].

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]. 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. "APManager"

With this tool it is possible to define a certain number of measures or parameters by combining articulatory trajectories given by Track, relative to specific reference points, lines or planes opportunely defined (see Figure 5). APmanager allows also to visualize and modify the patterns of these choosen parameters together with their relative velocity and accelleration and also to extract minimum and maximum points needed to identify and better specify articulatory gestures.

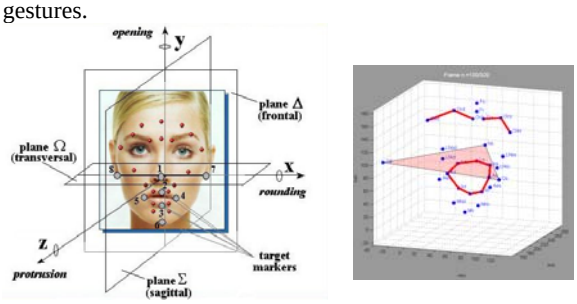

Figure 5: *Reference points and planes used for recording articulatory movements*.

2.4. "TXT/XMLediting"

This is an emotion specific XML editor explicitly designed for emotional tagged text. The APML mark up language [15] for behavior specification permits to specify how to markup the verbal part of a dialog move so as to add to it the "meanings" that the graphical and the speech generation components of an animated agent need to produce the required expressions. So far, the language defines the components that may be useful to drive a face animation through the facial description language (FAP) and facial display functions. The extension of such language is intended to support voice specific controls. An extended version of the APML language has been included in the FESTIVAL speech synthesis environment, allowing the automatic generation of the extended .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>, <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.

2.5. "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.

2.6. “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 [11] 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.

2.7. “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 (inspired by [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 6.

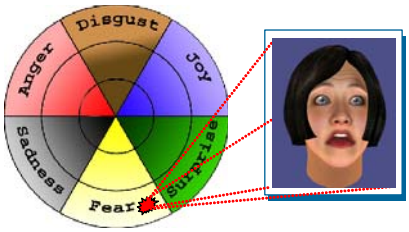

Figure 6: *Emotion Player. Clicking on 3-level intensity (low, mid, high) emotional disc [13], an emotional configuration (i.e. high -fear) is activated.*

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. New visualization tools have been introduced and new 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 European Project IST- 2001-37599 (http://pfstar.itc.it).

5. References

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

[2] Tisato G., Cosi P., Drioli C., Tesser F. (2005), “INTERFACE: a New Tool for Building Emotive/Expressive Talking Heads”, in Proc. INTERSPEECH 2005, Lisbon, Portugal, 781-784.

[3] Ferrigno G., Pedotti A. (1985), “ELITE: A Digital Dedicated Hardware System for Movement Analysis via Real-Time TV Signal Processing”, IEEE Trans. on Biomedical Engineering, BME-32, 943-950.

[4] Cosi P., Fusaro A., Tisato G. (2003), “LUCIA a New Italian Talking-Head Based on a Modified Cohen-Massaro’s Labial Coarticulation Model”, Proc. Eurospeech 2003, Geneva, Switzerland, 127-132.

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

[6] Ekman P. and Friesen W. (1978), Facial Action Coding System, Consulting Psychologist Press Inc., USA.

[7] Dutoit T. and Leich H. (1993), “MBR-PSOLA: Text-To-Speech Synthesis Based on an MBE re-Synthesis of the Segments Database”, Speech Commuication, vol. 13, no. 3-4, 167–184.

[8] Sommavilla G., Cosi P., Drioli C., Paci G. (2007), “SMS-FESTIVAL: a New TTS Framework”, Proc. MAVEBA 2007, Florence, (to be printed).

[9] Cohen M., Massaro D. (1993), “Modeling Coarticulation in Synthetic Visual Speech”, in Magnenat-Thalmann N., Thalmann D. (Eds), Models and Techniques in Computer Animation, Springer Verlag, 139-156.

[10] Tiede, M.K., Vatikiotis-Bateson, E., Hoole, P. and Yehia, H (1999), “Magnetometer data acquisition and analysis software for speech production research”, ATR Technical Report TRH 1999, ATR Human Information Processing Labs, Japan.

[11] Cosi P. and Hosom J.P. (2000), “Performance ‘General Purpose’ Phonetic Recognition for Italian”, Proc. of ICSLP 2000, Beijing, Cina, Vol. II, pp. 527-530, 2000.

[12] Pelachaud C., Magno Caldognetto E., Zmarich C., Cosi P. (2001), “Modelling an Italian Talking Head”, Proc. AVSP 2001, Aalborg, Denmark, September 7-9, 72-77.

[13] Cosi P., Magno Caldognetto E., Perin G., Zmarich C. (2000), “Labial Coarticulation Modeling for Realistic Facial Animation”, Proc. ICMI 2002, Pittsburgh, PA, USA, 505-510.

[14] Perin G. (2000-2001), Facce parlanti: sviluppo di un modello coarticolatorio labiale per un sistema di sintesi bimodale, Thesis, Univ. of Padova, Italy.

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

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