INTERFACE: a Matlab® Tools for Building Animated MPEG4 Talking Heads from Motion-Captured Data

Graziano Tisato, Piero Cosi, Carlo Drioli, Fabio Tesser
Istituto di Scienze e Tecnologie della Cognizione
Sezione di Padova “Fonetica e Dialettologia”
Via G. Anghinoni, 10 – 35121 Padova (ITALY)
+39 049 8274413
[tisato, cosi, drioli, tesser]@pd.istc.cnr.it

ABSTRACT
INTERFACE is an integrated software, designed and implemented in Matlab©, for building emotive/expressive talking heads from motion-captured data. INTERFACE simplifies and automates many of the operation 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 synthesis by means of an appropriate APML/VSML tagged language.

Categories and Subject Descriptors
I.6.7 [SIMULATION AND MODELING]: Simulation Support Systems

General Terms
Algorithms, Design, Human Factors, Standardization

Keywords
Talking Head, Facial Animation, Motion Capture, MPEG4.

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 man-machine interaction. 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.

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 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 [2], 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.

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:

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;
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 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.

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:

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

where \( F(n) \) is generated by a modified version of the Cohen-Massaro coarticulation model [6] as introduced in [9-10]. 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 3). 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 mediumlevel 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 3: APML/VSML mark-up language extensions for emotive audio/visual synthesis.](image)

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.

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 4.

![EmotionPlayer](image)

Figure 4: 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. ACKNOWLEDGMENTS

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

5. REFERENCES