# EMOTIONPLAYER: DALLA TEORIA ALLA PRATICA

*Piero Cosi, Carlo Drioli, Andrea Fusaro, Fabio Tesser, Graziano Tisato*
ISTC CNR – Sezione di Padova "Fonetica e Dialettologia" – Padova

## 1- Introduzione

In una situazione comunicativa di e-learning è di fondamentale utilizzo l'uso di strumenti che favoriscono la collaborazione e lo scambio di informazione tra i discenti e tra i discenti e i docenti.

In particolare nella didattica online la rete è utilizzata essenzialmente per l'erogazione di materiale didattico multimediale sia da parte del docente che da parte degli studenti (apprendimento collaborativo) e per la comunicazione nelle comunità di apprendimento. L'interazione dialogica può essere *asincrona* (e-mail, forum, newsletter) o *sincrona* (chat, audioconferenza, videoconferenza) e può variare inoltre nel contenuto e nella forma in base alla tipologia degli utenti: la comunicazione può infatti essere tra studente-docente, studente-tutor, tutor-docente, studente-studente (Anderson et al. 2002).

In particolare, poiché il sistema si basa sulla Comunicazione Mediata da Computer (*CMC*) (Baracco 2002), gli utenti trovano difficoltà nell'esprimere nei messaggi scritti gli aspetti interpersonali affettivi ed emotivi, in particolare all'interno delle aree d'interazione della chat e del forum.

Un contributo a queste problematiche è offerto dall'ISTC sez. di Padova all'interno del progetto PF-STAR *(Preparing future multisensorial interaction researc)*, per la messa a punto di una interfaccia uomo-macchina bimodale, cioè una Faccia Parlante in grado di sintetizzare vocalmente e visivamente (usando i corretti movimenti labiali e la corretta visual prosody) un testo scritto aggiungendo eventualmente le corrette emozioni.

## 2. La faccia parlante LUCIA e lo standard MPEG-4

Presso l'ISTC sez. di Padova da anni si sta sviluppando ***Lucia*** (Cosi et alii, 2003), una Faccia Parlante in italiano (Figura 1), basata su un sistema di sintesi bimodale da testo (Figura 2).

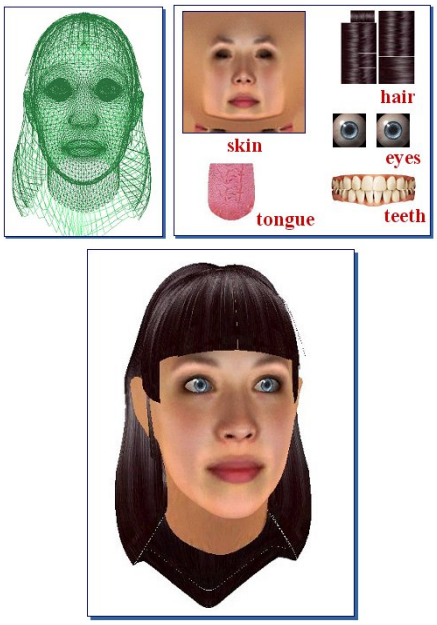


**Figura 1. La Faccia Parlante LUCIA**

*Lucia* parla in italiano mediante la versione italiana di FESTIVAL (Cosi et alii, 2001), la cui architettura è schematicamente illustrata in Figura 2. La Faccia Parlante è basata sullo standard MPEG-4 (MPEG www page) e su uno specifico modello di coarticolazione (Cohen & Massaro, 1993) appositamente sviluppato per rendere più fluidi e naturali i movimenti delle labbra.

*Lucia* è visualizzata in tempo reale sullo schermo e sincronizzata con il corrispondente segnale vocale fornito dal sistema di sintesi da testo. La sua animazione risulta molto fluida grazie ad una distribuzione ottimale dei poligoni e prevede la possibilità di essere utilizzata in remoto in applicazioni di tipo chat.

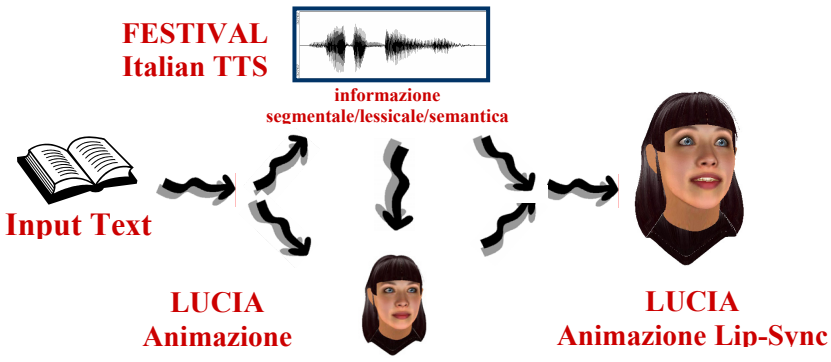


**Figura 2. Diagramma a blocchi dell'architettura di LUCIA**

### MPEG4 Animation

In MPEG-4 [9], gli *FDPs (Facial Definition Parameters)* definiscono la forma del modello mentre i *FAPs* (*Facial Animation Parameters*), definscono i movimenti facciali. Dato il modello,

la sua animazione è ottenuta tramite uno specifico *FAP-strem* che defisce i valori dei FAPs per ogni frame (figura 2). In ogni *FAP-stream* ogni frame ha due linee di parametri. Nella prima è indicata l'attivazione del particolare marker (valore 0 o 1), mentre nella seconda sono memorizzati, in termini di differenza dai precedenti,i valori target.

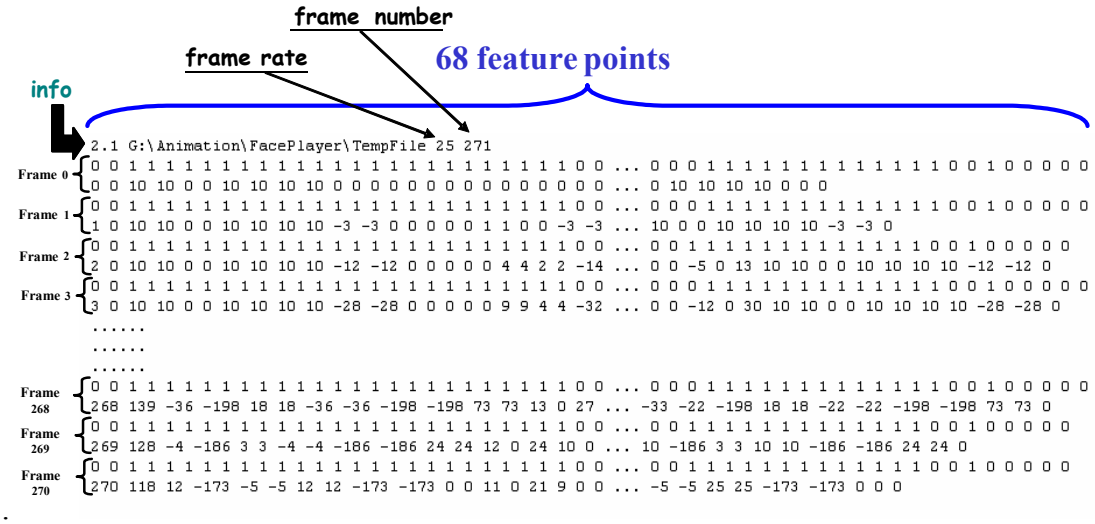


**Figura 3. Esempio di struttura di un FAP stream**

Nel nostro caso il modello faccciale utilizza un'approccio pseudo muscolare nel quale le contrazioni sono ottenute attraverso la deformzione della mesh poligonale attorno a punti chiave che corrispondono all'attaccatura dei muscoli facciali. Ogni feature point segue le specifiche MPEG4 dove ad ogni FAP corrisponde una minima azione facciale. Quando un FAP è attivo (per esempio quando l'inensità è non nulla) il corrispondente feature point si muove con l'intensità e la direzione indicati dallo stesso.

Utilizzando un'approccio pseudo muscolare, vengono deformati i punto della mesh poligonale che cadono all'interno della regione del feature point. L'espressione facciale è caratterizzata non solo dalla contrazione muscolare ma anche da una intensità e da una durata. L'intensità è ottenuta specificando il valore del fap mentre il fattore temporale è modellato tramite tre parametri chiamati: onset, apex e l'offset [14].

Il *FAP-stream* necessario per animare il FAE *(Facial Animation Engine)* può essere sintetizzato con l'utilizzo di uno specifico modello, come nel caso di LUCIA, o può essere ricostruito sulla base di dati ottenuti da sistemi optoelettronici come ELITE come avviene nel caso del software EmotionalPlayer.

Le configurazioni dei parametri facciali utilizzate nelle varie emozioni sono state estratte da un corpus di parlato emotivo raccolto all'interno del progetto europeo PF-Star, mediante il quale è

stato possibile analizzare in dettaglio alcune delle caratteristiche visive ed acustiche corrispondenti alle emozioni sopra elencate.

## 3 – EMOTIONALPLAYER

Alla data odierna la configurazione delle emozioni visuali sono progettate e raffinate, per mezzo di controllo visivo dei dati reali, con un software chiamato **EMOTIONALPLAYER,** progettato e realizzato in Matlab© sulla base dell'EmotionalPlayer (EP) e liberamente ispirato dal software Emotion disc. In futuro l'Emotional Player potrà gestire singoli movimenti facciali di una faccia sintetica progettata in MPEG-4 per creare dei rendering emozionali ed espressivi in Lucia.

I paramtri di onset e offset prima descritti rappresentano, rispettivamente, il tempo per cui l'espressione facciale appare e scompare; l'apex corrisponde alla durata per la quale l'espressione facciale è al massimo dell'intensità. Questi parametri sono fondamentali per caratterizzare le espressioni facciali.

Nel nostro sistema ogni espressione faciale è caratterizzata da un set di FAP. Ogni set di FAP permette per esempio la creazione delle 6 espressioni facciali corrispondenti alle 6 emozioni primarie di base di Ekman's (table2), scelte qui per semplicità, e per ogni espressione vengono simulati solo 3 livelli d'intensità (low, medium, high) corrispondenti alle zone concentriche del cerchio.

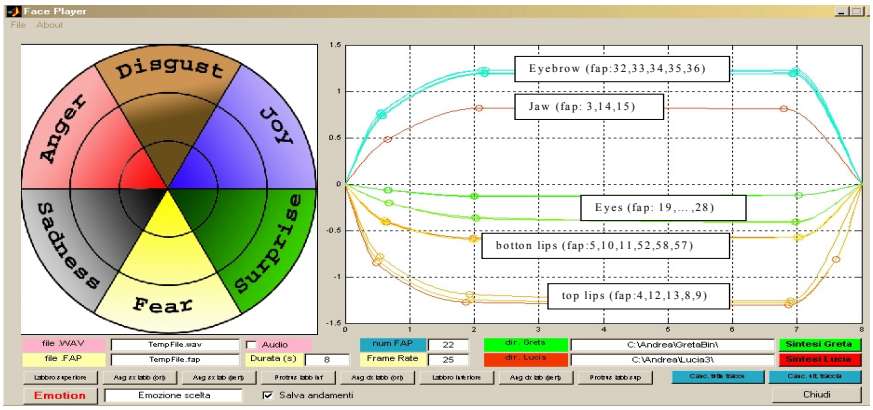


**Figura 4. EMOTIONALPLAYER**

.

| 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. 1. Le 6 emozioni primarie di Ekman's con le corrispondenti espressioni facciali**.**

Nel nostro sistema si distinguono le "Emozioni base" *EB(t)* dalle "emozioni rappresentate" *ED(t)*. Ci sono più funzioni al tempo *t.* Ogni *EB(t)* riguarda una speifica zona del volto come le sopracciglia, il mento, la bocca, le palpebre e così via. Le funzioni *EB(t)* incudono anche i movimenti facciali come i cenni del capo e i movimenti degli occhi. Ogni *EB(t)* è definita da un set di FAP MPEG-4:

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

Dove $v_1(t),\ldots,v_k(t)$ specificano le intensità dei valori dei FAP create dall'utente. Un *EB(t)* può inoltre essere definita da una combinazione di *EB'(t)* tramite l'operatore '+' nel seguente modo:

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

L'emozione rappresentata è infine ottenuta tramite una combinazione lineare:

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

Dove *EB(t)* è un'espressione facciale di base e 'c' è una costante. L'operatore '*' moltiplica ogni funzione EB(t) per una costante *'c'*. I valori di onset, offset e apex (per esempio la durata dell'espressione) dell'emozione sono determinati da una somma pesata di funzioni $v_k(t)$ ($k$ = 3,...,68) create tramite l'utilizo del mouse. Nella figura 5 sono evidenziati due semplici esempi di rabbia e felicità.

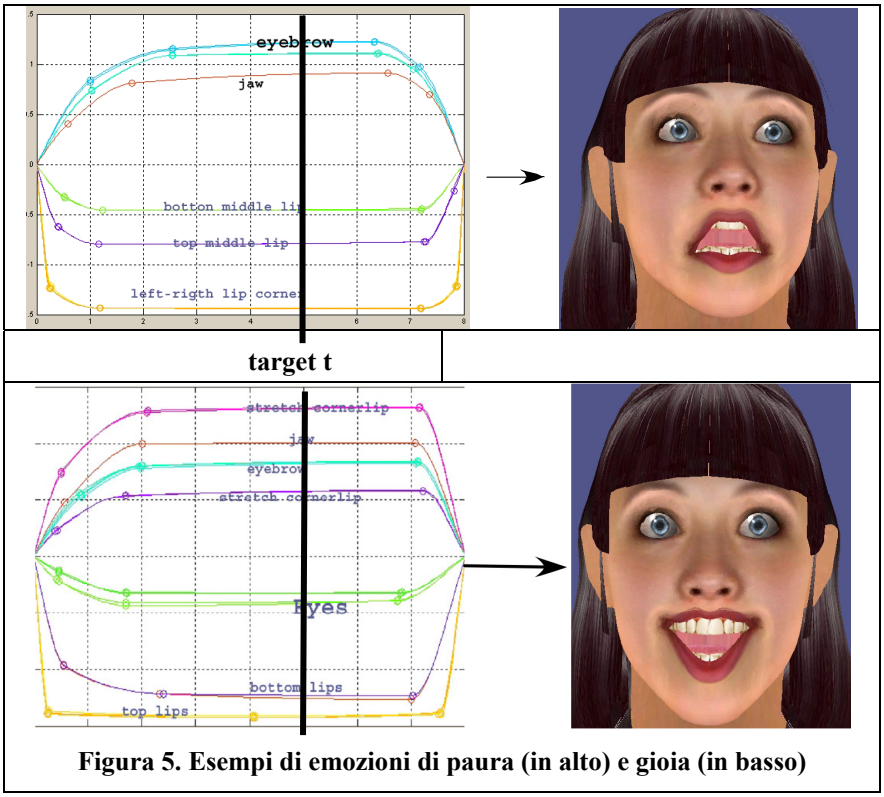


**Figura 5. Esempi di emozioni di paura (in alto) e gioia (in basso)**

### 3.1 XML_PLAYER software: evoluzione dell'EMOTIONALPLAYER

E' in corso d'implementazione il software *XML_PLAYER* che permette sia di interagire con la Faccia Parlante *Lucia* per la creazione di singole espressioni facciali rappresentanti le 6 emozioni di base (gioia, tristezza, rabbia, paura, disgusto, sorpresa) sia di generare un parlato emotivo coprodotto con movimenti facciali e labiali, rendendo così l'interazione più naturale, robusta ed amichevole.

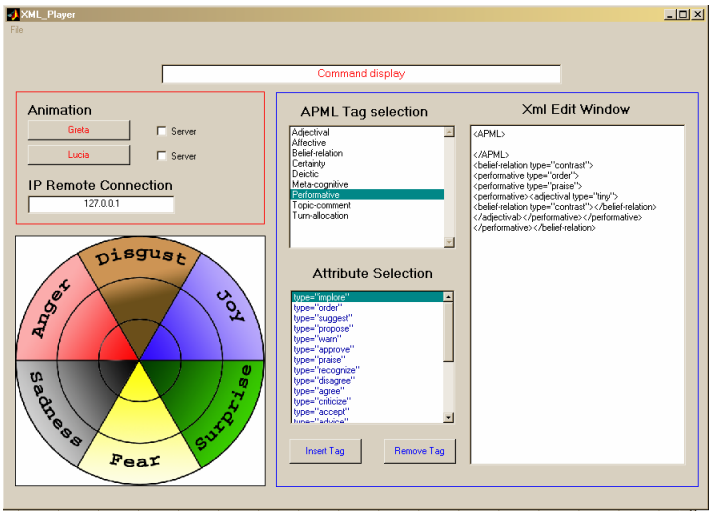


**Figura 6. Schermata del software XML_Player**

Il software si compone principalmente di due moduli.

- Un'interfaccia messa a punto all'ISTC sez. di Padova per realizzare un file di testo in linguaggio **APML** (*Affective Presentation Markup Language*) che permette di rappresentare le funzioni comunicative potenzialmente incluse in conversazioni naturali focalizzandosi sulla struttura della comunicazione e in particolare sul ruolo dei performativi come unità di base (De Carolis et al., 2002).
- L'**Emotion Disk** dove sono rappresentate le 6 emozioni di base con tre diversi livelli d'intensità (low, medium, high) per un totale di 18 stati emotivi. L'utente, selezionando un particolare livello d'intensità, otterrà l'animazione della Faccia Parlante sul computer remoto il cui indirizzo IP è specificato nella finestra "IP Remote Connection".

**Emotion Disk**

Prendendo spunto dall' "Emotion Disk" (Ruttkay et al. 2003), in EmotionPlayer sono rappresentate le 6 emozioni di base con tre diversi livelli d'intensità (basso, medio, alto) per un totale di 18 stati emotivi che possono essere selezionati dall'utente per ottenere in tempo reale la corrispondente animazione in LUCIA.

Come già detto nel nostro sistema ogni espressione facciale di base è caratterizzata da un particolare set di FAP (che coinvolge parti diverse della faccia).

Ciascuno dei tre livelli d'intensità è poi ottenuto moltiplicando il set di FAP che caratterizza l'emozione di base per una costante reale. Il file d'informazione così creato contenente l'emozione da riprodurre viene inviato tramite protocollo TCP/IP (***T**ransmission **C**ontrol **P**rotocol/**I**nternet **P**rotocol)* alla Faccia Parlante posta su un pc remota.

**Figura 7. Esempi di utilizzo dell'Emotion Disk**

## 3. Conclusioni

Nel testo scritto tutte le parole devono essere lette e interpretate, un processo più lento in percezione e meno spontaneo in produzione. L'utilizzo di un'informazione uditiva o visiva o bimodale uditivo-visiva, quale quella fornita dalla Faccia Parlante *Lucia*, potrebbe rendere più veloce la comprensione e l'interazione tra gli utenti delle chat didattiche. Infatti, mettendosi dal punto di vista dell'emittente, il testo scritto può non essere sempre soddisfacente in quanto di difficile comprensione. Per disambiguare la comunicazione negli ambienti di apprendimento viene proposta la tecnologia della Faccia Parlante espressiva, in grado di trasmettere parlato emotivo co-prodotto con movimenti facciali e labiali, corrispondenti a emozioni e atteggiamenti, rendendo così l'interazione più naturale, robusta ed amichevole e migliorando l'accessibilità all'e-learning.

## BIBLIOGRAFIA


Anderson L. e Ciliberti A., Monologicità e di(a)logicità nella comunicazione accademica, in C. Bazzanella (a cura di), Sul Dialogo. Contesti e forme di interazione verbale, Edizioni Angelo Guerini e Ass., Milano, 2002, pp. 91-105.

Baracco A., La comunicazione mediata dal computer, in C. Bazzanella (a cura di), Sul Dialogo. Contesti e forme di interazione verbale, Edizioni Angelo Guerini e Ass., Milano, 2002, pp. 253-267.

Bazzanella C., Prototipo, dialogo e configurazione complessiva, in C. Bazzanella (a cura di), Sul Dialogo. Contesti e forme di interazione verbale, Edizioni Angelo Guerini e Ass., Milano, 2002, pp. 19-44.

Cosenza A., I messaggi SMS, in C. Bazzanella (a cura di), Sul Dialogo. Contesti e forme di interazione verbale, Edizioni Angelo Guerini e Ass., Milano, 2002, pp. 193-207.

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.

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.

De Carolis, B., V. Carofiglio, M. Bilvi, C. Pelachaud (2002), APML, a Mark-up Language for Believable Behavior Generation. In: Proc. of AAMAS Workshop 'Embodied Conversational Agents: Let's Specify and Compare Them!', Bologna, Italy, July 2002

Ursini F., Multimodalità nella scrittura? Gli SMS tra telefoni cellulari, in Atti delle XI Giornate di Studio del Gruppo di Fonetica Sperimentale, "Multimodalità e multimedialità nella comunicazione", a cura di E. Magno Caldognetto e P. Cosi, (Padova 29/11-1/12 2000), UNIPRESS, Padova, pp. 75-80.

MPEG-4 standard. Home page: http://mpeg.telecomitalialab.com/standards/MPEG4.

World Wide Web Consortium (W3C) Home page: http://www.w3.org/XML/