GRETA E LUCIA:
DUE REALISTICHE FACCE PARLANTI ANIMATE MEDIANTE UN NUOVO MODELLO DI COARTICOLAZIONE

Piero Cosi*, Vincenzo Ferrari*, Emanuela Magno*,
Giulio Perin**, Graziano Tisato*, Claudio Zmarich*

*ISTC-SPFD CNR - Istituto di Scienza e Tecnologie della Cognizione
Sezione di Padova "Fonetica e Dialettologia" - Consiglio Nazionale delle Ricerche
e-mail: cosi@csrf.pd.cnr.it http://www.csrf.pd.cnr.it/

**UNIPD-DEI - Università di Padova - Dipartimento di Elettronica e Informatica
Via Gradenigo 6/a, 35131 Padova, ITALY
e-mail: giulioperin@yahoo.it

1. SOMMARIO

Per controllare automaticamente una faccia sintetica parlante, si sono recentemente imposti all'attenzione dei ricercatori i metodi basati sulla sintesi audiovisiva comandata direttamente da testo scritto, in cui il segnale acustico viene generato da un sistema di sintesi vocale (*TTS - Text-To-Speech synthesis*) e l'informazione fonetica estratta dal testo viene utilizzata per definire i corrispondenti movimenti articolatori.

Per la generazione di facce parlanti naturali, espressive e realistiche è necessario riprodurre fedelmente la variabilità contestuale dovuta alla reciproca influenza dei movimenti articolatori durante la produzione del segnale verbale ("coarticolazione").

In questo lavoro, viene illustrata una versione modificata del modello di coarticolazione proposto da Cohen e Massaro dove le caratteristiche dinamiche del modello sono state individuate mediante una tecnica semi-sutomatica di minimizzazione basata sui dati cinematici reali di specifici movimenti articolatori al fine di riprodurre più fedelmente i reali movimenti labiali coinvolti nella produzione del segnale verbale.

Il modello è stato applicato con successo a GRETA e, più recentemente, a LUCIA, due facce parlanti in italiano, di cui vengono illustrate le principali caratteristiche.

2. INTRODUZIONE

Ci sono molti modi per controllare una faccia parlante sintetica. Fra questi la parametrizzazione geometrica [1-2], il "*morphing*" fra specifiche configurazioni visive corrispondenti al segnale verbale [3], i modelli basati sulla simulazione del funzionamento dei muscoli o "pseudo" muscoli facciali [4-5], sembrano i più attraenti. Recentemente, vi è inoltre un elevato interesse per i sistemi di sintesi audiovisivi da testo scritto [6-7], in cui il segnale acustico è generato da un sistema di sintesi da testo scritto (*TTS*) e l'informazione fonetica estratta dal testo viene utilizzata per definire e comandare i movimenti articolatori riprodotti dall'agente parlante.

Per la generazione di animazioni facciali realistiche è necessario riprodurre la variabilità contestuale causata dalla reciproca influenza dei movimenti articolatori nella produzione del segnale verbale. Questo fenomeno definito con il termine coarticolazione [8], è veramente complesso e difficile da modellare. Sono state studiate moltissime strategie di

coarticolazione, che possono anche differire in lingue diverse [9], ed in questo lavoro viene illustrata una versione modificata [10] del modello di coarticolazione proposto da Cohen e Massaro [11] basato sulla "*gestural theory of speech production*" di Löfqvist [12]. Per determinare le caratteristiche dinamiche del modello, è stata utilizzata una tecnica semi-automatica di minimizzazione basata sui dati cinematici reali di specifici movimenti articolatori labiali acquisiti da un sistema opto-elettronico, denominato ELITE [13], al fine di riprodurre più fedelmente i movimenti labiali coinvolti nella produzione del segnale verbale.

Questo modello è stato applicato con successo a GRETA [14-15] e, più recentemente, a LUCIA [16], due facce parlanti in italiano, mediante la versione italiana del sistema di sintesi da testo scritto denominata FESTIVAL [17], entrambe basate sullo standard di animazione facciale denominato MPEG-4 [18][1].

3. IL MODELLO DI COARTICOLAZIONE

Come accennato nell'introduzione, il modello di coarticolazione proposto da Cohen e Massaro [11] è fortemente ispirato dalla "*gestural theory of speech production*" di Löfqvist [13]. Ogni fonema è specificato in termini di parametri articolatori di controllo (arrotondamento delle labbra, abbassamento e innalzamento del labbro superiore e inferiore, protrusione delle labbra) caratterizzati da un valore "*target*" da raggiungere e da una funzione di dominanza. Le funzioni di dominanza associate ad ogni singolo fonema successivo si sovrappongono nel tempo e specificano il grado di influenza che un segmento vocale ha sugli articolatori coinvolti nella produzione dei segmenti precedenti e successivi.

La traiettoria articolatoria finale di uno specifico parametro è data quindi dalla somma della media pesata di tutte le dominanze scalate sulla base del modulo dei corrispondenti valori *target* di ogni singolo segmento. Per una sequenza di N fonemi, se T_i è l'ampiezza dell'i-esimo *target*, t_i la sua posizione temporale e $D_i(t)$ la sua dominanza associata, la funzione finale di un parametro è data da:

$$F(t) = \frac{\sum_{i=1}^{N} T_i \cdot D_i(t - t_i)}{\sum_{i=1}^{N} D_i(t - t_i)} \qquad (1)$$

dove la dominanza, avente la forma di una funzione esponenziale negativo, è data da:

$$D(\tau) = \begin{cases} \alpha e^{-\theta_{\mathrm{bw}} |\tau|^c} & \text{if } \tau \le 0 \\ \alpha e^{-\theta_{\mathrm{fw}} |\tau|^c} & \text{if } \tau > 0 \end{cases} \qquad (2)$$

dove α indica il modulo delle dominanza, θ_{bw} e θ_{fw} rappresentano le sue estensioni temporali all'indietro (*bw - backward*) e in avanti (*fw - forward*) e la potenza c ne influenza il grado di attivazione. L'influenza o dominanza di un segmento su quelli a lui vicini prima

[1] Parte di questo lavoro è stato possibile grazie alle attività sviluppate nell'ambito dei progetti: MPIRO (Multilingual Personalized Information Objects, European Project IST-1999-10982, http://www.ltg.ed.ac.uk/mpiro/), TICCA (Tecnologie cognitive per l'interazione e la cooperazione con agenti artificiali, progetto congiunto fra il CNR e la Provincia Autonoma Trentina), e PF-STAR (Preparing Future multiSensorial inTerAction Research, European Project IST- 2001-37599, http://pfstar.itc.it/).

aumenta e poi diminuisce raggiungendo il suo massimo nella posizione temporale del *target* articolatorio.

4. IL MODELLO DI COARTICOLAZIONE MODIFICATO

Il metodo implementato da Cohen e Massaro può essere migliorato per realizzare una descrizione più accurata delle transizioni fra *target* articolatori successivi, soprattutto a differenti valori di velocità di eloquio, e per risolvere parecchie difficoltà incontrate nella modellizzazione articolatoria delle consonanti bilabiali e labiodentali. Questo obiettivo è stato raggiunto adottando una nuova versione più generale delle funzioni di dominanza ottenuta aggiungendo alcune componenti di *resistenza temporale* e *forma* al modello originale. Nel modello originale, infatti, il parametro c è impostato ad un valore costante unitario, ma in un contesto più generale può essere differente per ogni fonema e può anche assumere valori differenti nel caso *backward* e *forward* (c_{bw}, c_{fw}). Da questo punto di vista può essere interpretato come il tasso di attivazione e di rilascio del gesto articolatorio. Le variazioni di c_{bw} and c_{fw} generano comportamenti qualitativamente differenti delle dominanze, e conseguentemente del corrispondente parametro di controllo, che risultano particolarmente evidenti ad elevate velocità di eloquio [10, fig. 2]

Come riportato in [12], l'utilizzazione di diversi valori di dominanza risente dell'idea, presente nella teoria di Bladon e di Al-Bamerny" [9], di utilizzare un coefficiente numerico per rappresentare in ogni segmento la resistenza coarticolatoria associata ad alcune caratteristiche fonetiche. Questo è strettamente collegato all'idea di utilizzare differenti valori di dominanza per ogni fonema e si riflette ad esempio sulle modalità utilizzate dalle labbra per raggiungere la loro posizione finale. Ad un'elevata velocità di eloquio le dominanze sono molto vicine fra loro e anche se le loro ampiezze sono elevate la traiettoria finale non riesce a far raggiungere i valori *target*. Questo costituisce un problema in tutti quei casi in cui il *target* articolatorio deve essere obbligatoriamente raggiunto come nel caso della produzione delle occlusive bilabiali (/p, b, m/) e delle fricative labiodentali (/f, v/). Per ovviare a questo problema è stato introdotto il concetto di resistenza coarticolatoria che è stato applicato all'estensione temporale delle dominanze.

Ad ogni dominanza è stato associato un esponenziale negativo $R(\tau)$, denominato funzione temporale di resistenza [10, formula (4)], la cui principale caratteristica è rappresentata dal fatto che la sua estensione temporale in avanti e all'indietro può essere diversa a seconda del valore del *coefficiente di resistenza* k_{R} dei fonemi precedenti e successivi. In altre parole, se la resistenza del fonema successivo è massima ($k_{\mathrm{R}} = 1$), l'estensione temporale in avanti della funzione di resistenza è uguale alla distanza temporale fra il fonema *target* corrente e quello successivo. In questo modo la combinazione delle funzioni di dominanza e resistenza $D(\tau) \cdot R(\tau)$ raggiunge il valore nullo in corrispondenza dell'istante in cui la dominanza del fonema successivo raggiunge il suo valore massimo, di conseguenza il *target* articolatorio può essere forzatamente raggiunto. Per $k_{\mathrm{R}} < 1$, l'estensione di $R(\tau)$ aumenta seguendo una procedura ricorsiva definita in [10].

E' stata introdotta inoltre una funzione *forma* [10, formula (5)] al fine di modellare più efficacemente il comportamento articolatorio in prossimità dei *target*. Questa funzione è

utile quando si vogliono descrivere alcuni specifici andamenti dei parametri articolatori come ad esempio una particolare pendenza in prossimità dei *target* o una particolare transizione caratterizzata da un'iniziale forte caduta seguita da una più dolce. In conclusione, la funzione parametro originale (1) è stata modificata per includere le nuove funzioni di resistenza temporale $R(\cdot)$ e di forma (*shape*) $S(\cdot)$, precedentemente definite, ottenendo così :

$$F_{new}(t) = \frac{\sum_{i=1}^{N} T_i \cdot S_i(t-t_i) \cdot R_i(t-t_i) \cdot D_i(t-t_i)}{\sum_{i=1}^{N} R_i(t-t_i) \cdot D_i(t-t_i)} \qquad (3)$$

dove la funzione di resistenza temporale è stata inclusa anche a denominatore a causa della sua stretta relazione con le dominanze ad essa associate.

4. I DATI DI ANALISI

I valori dei coefficienti del nuovo modello sono stati determinati a partire da un corpus di movimenti ariticolatori prodotti da un soggetto italiano che ha pronunciato gli stimoli simmetrici VCV, in cui V è uno dei fonemi vocalici cardinali /i/, /a/ o /u/ mentre C è uno dei fonemi consonantici. In questo corpus sono contenute le traiettorie spazio-temporali di 6 parametri (apertura del labbro superiore, apertura del labbro inferiore, protrusione del labbro superiore, protrusione del labbro inferiore, arrotondamento delle labbra, apertura della mandibola) registrati dal sistema optoelettronico denominato ELITE [11]. La procedura di stima dei parametri è basata su metodo di minimizzazione dei minimi quadrati:

$$e(t) = \sum_{i=1}^{N} \left(Y(n) - F(n) \right)^2 \qquad (4)$$

fra i dati reali (*Y(n)*) e le curve ottenute in uscita dal modello (*F(n)*) per 5 ripetizioni dello stesso tipo di sequenze. E' stata utilizzata una procedura automatica di ottimizzazione caratterizzata da una forte proprietà di convergenza [14] anche se, a causa della presenza di parecchi minimi locali, è stata guidata manualmente al fine di evitare risultati indesiderati. Nella Figura 1 è illustrato un semplice esempio dell'andamento delle traiettorie articolatorie del parametro di apertura delle labbro inferiore. Come si può notare le curve reali e quelle simulate sono molto simili fra loro infatti in media l'errore fra le traiettorie reali e quelle simulate risulta essere inferiore a 0.3 mm.

5. GRETA e LUCIA

Il modello modificato [10] è stato applicato con successo a GRETA [14-15] e, più recentemente, a LUCIA [16], due agenti animati basati entrambi sullo standard MPEG-4 [18] e parlanti in italiano mediante la versione italiana di FESTIVAL [17], come graficamente illustrato nel diagramma a blocchi di Figura 2.

Come detto, "Greta" e "Lucia" sono dei motori di animazione facciale compatibili con lo standard MPEG-4 ed entrambe possono essere utilizzate per realizzare un *decoder* compatibile con il cosiddetto "*Predictable Facial Animation Object Profile*" [18]. Lo standard MPEG-4 specifica un insieme di parametri di animazione facciale "*Facial*

Animation Parameters" (FAP), corrispondenti a specifiche azioni di deformazione della forma a riposo del modello della faccia. Una particolare sequenza di animazione viene generata mediante successive deformazioni del modello facciale seguendo opportuni valori FAP che indicano rispettivamente l'intensità dell'azione deformante e la sua estensione temporale. Il modello poi viene visualizzato in tempo reale sullo schermo e sincronizzato con il corrispondente segnale vocale fornito, in questo caso dal sistema di sintesi da testo scritto per l'italiano FESTIVAL.

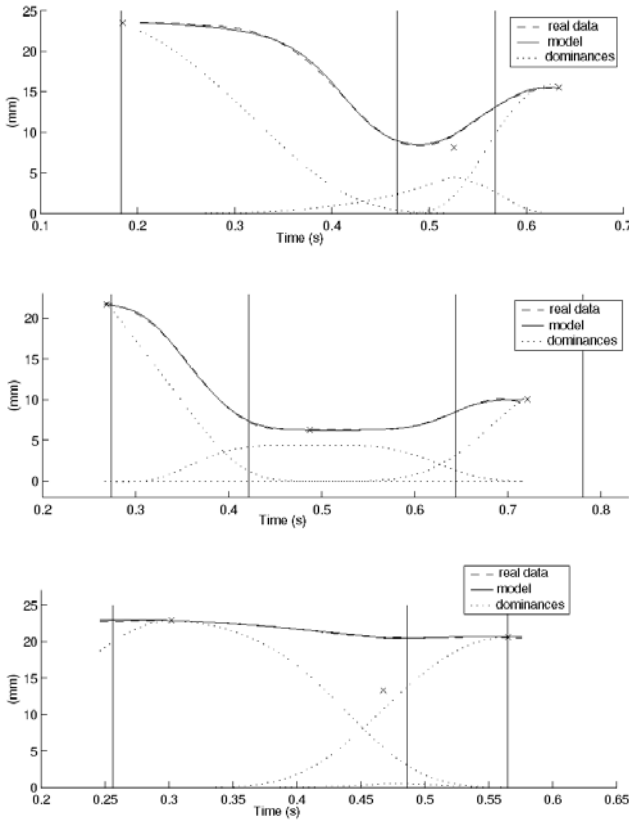

Figura 1. Traiettorie del parametro di apertura del labbro inferiore durante la produzione delle sequenze isolate /'a d a/ (a), /'a dz a/ (b) e /'a l a/ (c). Le linee tratteggiate indicano le funzioni di dominanza.

Sia Greta che Lucia emulano le funzionalità dei muscoli "mimici" facciali mediante l'utilizzo di specifiche "funzioni di deformazione" agenti in specifici punti del modello. L'attivazione di queste funzioni è determinata da specifici parametri che codificano le varie azioni muscolari sulla faccia e queste azioni possono quindi essere modificate a piacere per generare l'animazione desiderata. Questi parametri, in MPEG-4 denominati FAP, hanno un ruolo fondamentale nel rendere naturale il movimento del modello. L'azione muscolare è resa esplicita mediante la deformazione di un reticolo poligonale tridimensionale costruito attorno ad alcuni punti specifici definiti sulla faccia "*Facial*

Definition Parameters" (FDP), che corrispondono all'attaccatura sulla pelle dei muscoli mimici. Il movimento esclusivo degli FDP non è da solo sufficiente a muovere in modo omogeneo il modello 3D nella sua interezza. E' per questo, infatti, che ad ogni FDP è collegata una particolare "zona d'influenza", costituita da un'ellisse, contenente quelle zone della pelle i cui movimenti sono strettamente connessi. Dopo aver stabilito tutte le relazioni di corrispondenza per tutti gli FDP e tutti i vertici, i punti del modello 3D possono essere mossi simultaneamente ed in modo omogeneo utilizzando, per ogni FDP, una funzione di pesatura del movimento dei vertici collegati, caratterizzata dalla forma di un coseno-rialzato.

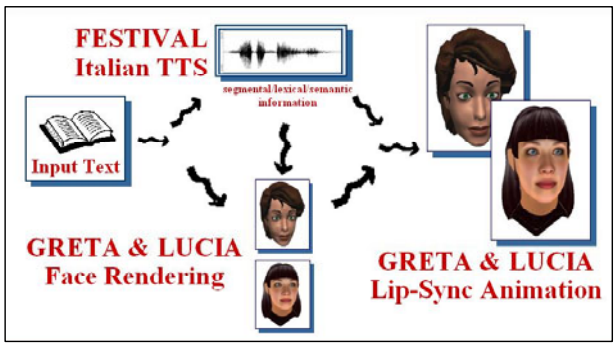

Figura 2: Greta and Lucia, due agenti parlanti in italiano

La differenza principale fra i due modelli risiede essenzialmente nell'utilizzo di *texture* reali per LUCIA (vedi Figura 3), in un diverso numero di poligoni utilizzati e nella possibilità, disponibile solo nel caso di LUCIA, di generare un modello poligonale 3D importando direttamente la sua struttura da un file VRML [17]. Attualmente sia Greta che Lucia sono due giovani facce 3D femminili e ad esempio LUCIA è costruita mediante 25423 poligoni, 14116 appartenenti alla pelle, 4616 ai capelli, 2688x2 agli occhi, 236 alla lingua e 1029 ai denti. Nel caso di LUCIA il modello è diviso in due parti principali: la pelle e gli articolatori interni (occhi, lingua, denti). Questa suddivisione è particolarmente utile per l'animazione, poiché soltanto la pelle è direttamente influenzata dall'azione dei pseudo-muscoli mimici facciali è costituisce quindi un elemento unitario, mentre le altri componenti anatomiche risultano essere indipendenti fra loro e si muovono in modo più rigido seguendo esclusivamente delle traslazioni e/o rotazioni (per esempio gli occhi ruotano su se stessi attorno al loro punto centrale). Utilizzando questa strategia i poligoni sono distribuiti in modo tale da rendere l'effetto visivo molto naturale evitando di visualizzare possibili "discontinuità" nel modello 3D soprattutto in fase di animazione.

6. OSSERVAZIONI CONCLUSIVE

Il modello modificato di coarticolazione di Cohen-Massaro riesce a descrivere con notevole precisione la cinematica dei parametri articolatori (RMSE < 0.3 mm) e riesce, inoltre, a ben rappresentare le consonanti bilabiali (/p, b, m/) e labiodentali (/f, v/), anche con velocità di eloquio elevate.

I motori di animazione facciale GRETA e LUCIA, pur essendo simili ad altri modelli basati sullo standard MPEG, mediante l'utilizzazione del nuovo modello di coarticolazione, risultano possedere un movimento assai più naturale.

La qualità generale dell'animazione dovrà essere analizzata, sia in GRETA che in LUCIA, mediante adeguati test percettivi.

In futuro verranno simulate le "emozioni", così come il movimento di altri importanti articolatori quali ad esempio la lingua per un'animazione più naturale e realistica.

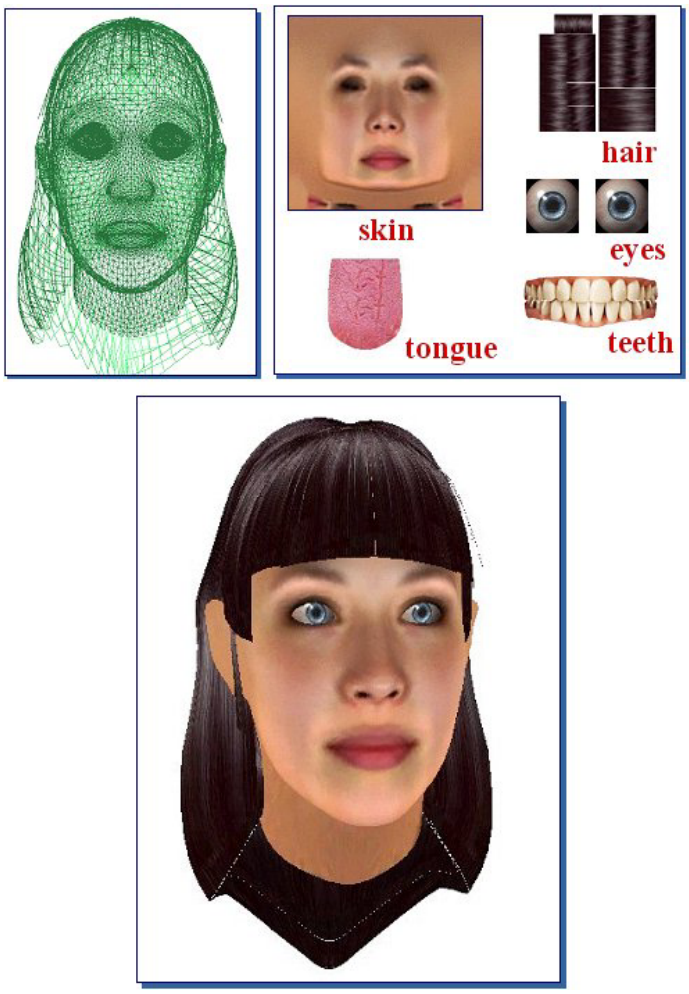

Figura 3: Wireframe e texture in Lucia

BIBLIOGRAFIA

[1] Massaro D.W., Cohen M.M., Beskow J., Cole R.A., "Developing and Evaluating Conversational Agents", in Cassell J., Sullivan J., Prevost S., Churchill E. (Editors), Embodied Conversational Agents, MIT Press, Cambridge, MA, 2000, pp. 287-318.

[2] Le Goff, B. Synthèse à partir du texte de visages 3D parlant français, PhD thesis, Grenoble, France, October 1997.

[3] Bregler C., Covell M., Slaney M., “Video Rewrite: Driving Visual Speech with Audio”, in Proc. of SIGGRAPH ’97, 1997, pp. 353-360.

[4] Lee Y., Terzopoulos D., Waters K., “Realistic Face Modeling for Animation”, in Proceedings of SIGGRAPH ’95, 1995, pp. 55–62.

[5] Vatikiotis-Bateson E., Munhall K.G., Hirayama M., Kasahara Y., Yehia H., “Physiology-Based Synthesis of Audiovisual Speech”, in Proc. of 4th Speech Production Seminar: Models and Data, 1996, pp. 241-244.

[6] Beskow J., “Rule-Based Visual Speech Synthesis,” in Proc. of Eurospeech ’95, , Madrid, September 1995.

[7] B. LeGoff and C. Benoit. (1996) A text-to-audiovisualspeech synthesizer for French. In Proc. of the ICSLP '96, Philadelphia, USA.

[8] Farnetani E., Recasens, “Coarticulation Models in Recent Speech Production Theories”, in Hardcastle W.J. (Editors), Coarticulation in Speech Production, Cambridge University Press, Cambridge, 1999.

[9] Bladon, R.A., Al-Bamerni, A., “Coarticulation resistance in English \l\”, Journal of Phonetics, 4, 1976, pp. 135-150.

[10] Cosi P., Perin G., “Labial Coarticulation Modeling for Realistic Facial Animation”, in Proc. of ICMI ‘02, October 14-16, 2002 Pittsburgh, PA, USA.

[11] 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.

[12] Löfqvist, A. “Speech as Audible Gestures”, in Hardcastle W.J., Marchal A. (Editors.), Speech Production and Speech Modeling, Dordrecht: Kluwer Academic Publishers, 1990, pp. 289-322.

[13] Ferrigno G., Pedotti A., "ELITE: A Digital Dedicated Hardware System for Movement Analysis via Real-Time TV Signal Processing", in IEEE Transactions on Biomedical Engineering, BME-32, 1985, pp. 943-950.

[14] Pasquariello S. (2000), Modello per l’animazione facciale in MPEG-4, M.S. thesis, University of Rome, 2000.

[15] Pelachaud C., Magno Caldognetto E., Zmarich C., Cosi P. (2001), “Modelling an Italian Talking Head”, in Proceedings of the International Conference on Auditory-Visual Speech Processing - AVSP’2001, Aalborg, Denmark, Settembre 7-9 2001, pp. 72-77.

[16] Cosi P., Ferrari V., Magno Caldognetto E., Perin G., Tisato G. and Zmarich C. (2003), “LUCIA a New Italian Talking-Head Based on a Modified Cohen-Masaro’s Labial Coarticulation Model”, (submitted to Eurospeech 2003).

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

[18] Mpeg-4 standard. Home page: http://mpeg.telecomitalialab.com/standards/mpeg4.