UN NUOVO MODELLO DI COARTICOLAZIONE PER L'ANIMAZIONE FACCIALE

Piero Cosi*, Vincenzo Ferrari*, Emanuela Magno Caldognetto*,
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 www: 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

SOMMARIO

Lo sviluppo di facce parlanti naturali, espressive e realistiche rende necessaria la riproduzione fedele della 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, in cui le caratteristiche dinamiche sono state individuate mediante una tecnica semi-automatica di minimizzazione basata sui dati cinematici reali di specifici movimenti articolatori al fine di riprodurre più fedelmente i 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 brevemente introdotte le principali caratteristiche.

1. INTRODUZIONE

Esistono molti metodi per controllare automaticamente una faccia sintetica parlante. Fra questi, quelli ritenuti in letteratura più interessanti sono senza dubbio i metodi a "parameterizzazione geometrica" (Massaro et alii, 2000; Le Goff, 1997), i metodi basati sul "morphing" fra differenti configurazioni articolatorie/visive" (Bregler et alii, 1997) e i metodi basati sui modelli fisiologici dei muscoli e pseudo-muscoli facciali (Lee et alii, 1995; Vatikiotis-Bateson et alii, 1996). Più recentemente, si sono imposti all'attenzione dei ricercatori anche i metodi basati sulla sintesi audiovisiva comandata direttamente da testo scritto (Beskow, 1995; LeGoff & Benoit (1996), 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 di sequenze fonetiche. Questo particolare fenomeno, definito "coarticolazione" (Farnetani & Recasens, 1999), è estremamente complesso e difficile da modellare. Esistono infatti numerevoli strategie coarticolatorie e queste possono anche differire in funzione della lingua utilizzata (Bladon & Al-Bamerni, 1976).

In questo lavoro, viene illustrata una versione modificata del modello di coarticolazione proposto da Cohen & Massaro (Cohen & Massaro, 1993) e basato sulla "*gestural theory of speech production*" di Löfqvist (Löfqvist, 1990).

Fortemente ispirati da Cohen et alii (Cohen et alii, 1998), "…Whatever the system, rather than tuning the control strategies by hand as has been done in the past, we need to use the mass of available static and dynamic observations of real humans to educate the systems to be more realistic and accurate..." ("…Indipendentemente dal sistema utilizzato, invece di determinare e regolare a mano le strategie di controllo, come è stato fatto in passato, abbiamo bisogno di utilizzare l'enorme mole di dati statici e dinamici relativi a reali osservazioni umane al fine di allenare i sistemi di animazione facciale ad essere più realistici ed accurati…"), per determinare le caratteristiche dinamiche del modello, è stata utilizzata una tecnica semi-sutomatica di minimizzazione basata sui dati cinematici reali di specifici movimenti articolatori labiali acquisiti da un sistema opto-elettronico, denominato ELITE (Ferrigno et alii, 1985), al fine di riprodurre più fedelmente i reali movimenti labiali coinvolti nella produzione vocale.

Il modello è stato applicato con successo a GRETA (Pasquariello, 2000; Pelachaud et alii, 2001) e più recentemente a LUCIA (Cosi, 2002, Cosi, 2003), due facce parlanti emotive ed espressive in italiano, di cui verranno successivamente illustrati alcuni esempi di animazione che dimostrano la naturalezza ottenuta nella simulazione dei movimenti articolatori facciali[1].

2. LA COARTICOLAZIONE

Nell'ambito del controllo labiale un aspetto fondamentale per la conformazione finale degli articolatori è dato dall'influenza del contesto in cui un particolare fonema si trova inserito. Tale fenomeno, comunemente definito come "coarticolazione", è stato ed è tutt'oggi oggetto di numerosi studi. I risultati ottenuti, talvolta contrastanti, dimostrano l'incredibile complessità del meccanismo di produzione del parlato e la difficoltà nel creare un modello che ne riproduca efficacemente il comportamento.

I sistemi basati sulla rappresentazione attraverso segmenti fonetici concatenati presentano due principali diffcoltà:

[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/).

- segmentare significa stabilire dei confini di demarcazione tra un fonema e l'altro, mentre il parlato è un processo continuo in cui non si possono determinare frontiere così nette
- un tale frazionamento suggerisce l'idea non corretta di indipendenza dell'unità fonetica (in realtà essa è fortemente dipendente dal contesto in cui si trova)

Il movimento dei differenti articolatori per la produzione di successivi fonemi si sovrappone e interagisce col movimento articolatorio dei segmenti adiacenti. Questo fa capire perchè gli studi sui fenomeni coarticolatori abbiano un grande rilievo. Da una parte si cercano delle teorie che ne spieghino l'origine, la natura ed il funzionamento, dall'altra si vuole creare dei modelli che ne predicano i dettagli. Gli studi coarticolatori riguardano due campi principali:

- la variabilità acustica
- la variabiltà articolatoria

Il tentativo di spiegare la variabilità degli aspetti coarticolatori ha dato vita a numerose teorie e modelli, di cui un breve elenco è illustrato in Tabella 1. Data la complessità dei comportamenti articolatori, gli studi da cui sono conseguite le varie teorie fanno in genere riferimento ad analisi di aspetti particolari e circoscritti

- la "variabilità adattativa" (Lindblom, 1966)
- il modello di Ohman (Ohman, 1966, 1967)
- la "fonologia generativa" (Chomsky & Halle 1968)
- la teoria dell'estensione delle caratteristiche (Daniloff & Moll, 1973) con particolare riferimento al modello look-ahead (Henke 1966, Kozhevnikov & Chistovich 1965)
- il modello della "resistenza alla coarticolazione" (Bladon & Al-Bamerni, 1976)
- il modello "a finestra" (Keating , 1988, 1990)
- il modello della "coproduzione gestuale" (Salzman & Munhall, 1989),(Munhall & Löfqvist, 1992)
- il modello time-locked (Bell-Berti & Harris, 1981)
- il modello ibrido (Al-Bamerni & Bladon 1982)
- il modello ad "espansione del movimento" (Abry & Lallouache, 1991)

Tabella1. Elenco di alcune delle teorie e dei modelli di coarticolazione apparsi in letteratura (vedi Farnetani & Recasens 1999)

3. IL MODELLO DI COARTICOLAZIONE DI COHEN-MASSARO

Uno dei modelli più interessanti apparsi in letteratura è sicuramente quello adottato da Cohen e Massaro (Cohen & Massaro, 1993). Essi si basarono sul modello gestuale di Löfqvist (Löfqvist, 1990), in cui ad ogni singolo gesto articolatorio è associata una funzione di dominanza con le stesse caratteristiche dei gesti fonetici. Un funzione di dominanza è caratterizzata da una propria ampiezza, durata, e grado di attivazione come indicato in Figura 1.

L'ampiezza determina l'importanza relativa del gesto per il segmento; la durata stabilisce l'estensione del movimento ed influisce sul grado di sovrapposizione che ne conseguirà; il grado di attivazione caratterizza il fatto che il gesto si avvia in modo più o meno graduale.

Per rappresentare una dominanza che caratterizzasse in modo differente la coarticolazione anticipatoria e perseverativa, è stata utilizzata una funzione esponenziale asimmetrica del tipo

$$D(\tau) = \begin{cases} \alpha e^{-\theta_{\mathrm{bw}}|\tau|^c} & \text{if } \tau \le 0 \\ \alpha e^{-\theta_{\mathrm{fw}}|\tau|^c} & \text{if } \tau > 0 \end{cases} \qquad (1)$$

in cui τ rappresenta la distanza temporale dal centro del segmento fonetico, α l'ampiezza della funzione di dominanza; θ_{bw} e θ_{fw} caratterizzano rispettivamente estensione anticipatoria ed estensione perseverativa ed infine *c* indica il grado di attivazione del gesto. Queste funzioni vengono combinate nel tempo con il valore stimato del target articolatorio e normalizzate per ottenere l'andamento complessivo

$$F(t) = \frac{\sum_{i=1}^{N} T_i \cdot D_i(t - t_i)}{\sum_{i=1}^{N} D_i(t - t_i)} \qquad (2)$$

dove *N* si riferisce al numero totale di segmenti, T_i e $D_i(t)$ rappresentano target e funzione di dominanza del segmento i-esimo centrato in t_i. In Figura 1 è illustrato un esempio dell'andamento, risultante dalla sovrapposizione di 3 fonemi adiacenti, del parametro rappresentativo dell'apertura labiale.

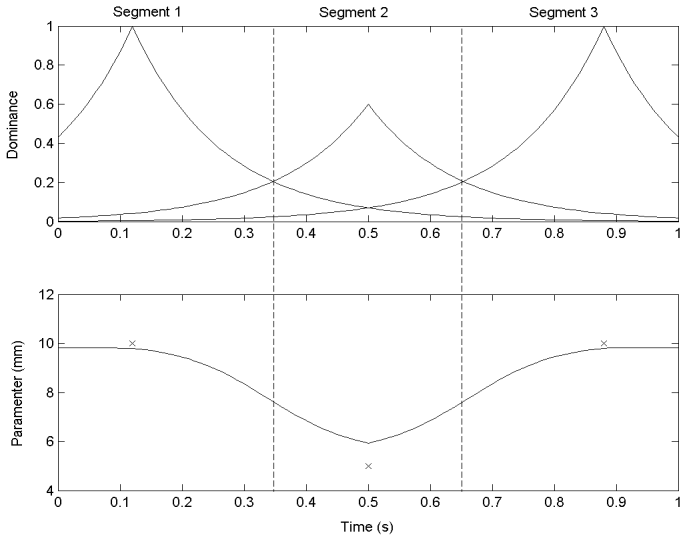


Figura 1. Funzioni di Dominanza di tre segmenti successivi (in alto) e andamento risultante di un parametro (in basso) secondo il modello di Cohen e Massaro (Cohen & Massaro, 1993).

4. IL MODELLO DI COARTICOLAZIONE MODIFICATO

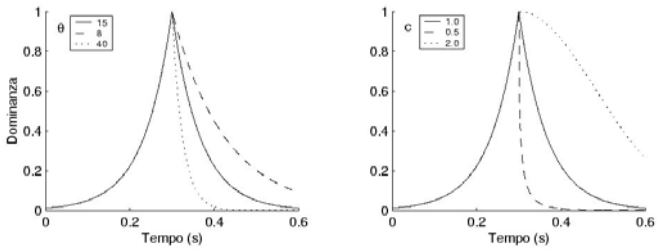
Il metodo implementato da Cohen e Massaro può essere migliorato per realizzare una descrizione più accurata delle transizioni fra *target* articolatori successivi, soprattutto a differenti valori di velocità di eloquio, e per risolvere parecchie difficoltà incontrate nella modellizzazione articolatoria delle consonanti bilabiali e labiodentali. Questo obiettivo è stato raggiunto adottando una nuova versione più generale delle funzioni di dominanza e aggiungendo al modello originale alcune componenti di *resistenza temporale* e *forma*.

4.1 La funzione di dominanza

Come precedentemente introdotto, tre coefficienti definiscono la conformazione di una funzione di dominanza, due per determinarne l'estensione temporale e uno che ne rappresenta il grado di attivazione. Nella versione originale del modello si utilizza un valore costante del coefficiente di attivazione. In base al confronto con andamenti cinematici reali si è visto che un valore di c pari a 1 fornisce i migliori risultati. È opportuno tuttavia considerare in un contesto più generale che il grado di attivazione del gesto articolatorio vari a seconda del fonema e che possa essere diverso per l'estensione anticipatoria e per quella perseverativa. Considereremo da ora in poi, quindi, una funzione di dominanza più generale del tipo

$$D(\tau) = \begin{cases} \alpha e^{-\theta_{\mathrm{bw}} |\tau|^{c_{\mathrm{bw}}}} & \text{if } \tau \le 0 \\ \alpha e^{-\theta_{\mathrm{fw}} |\tau|^{c_{\mathrm{fw}}}} & \text{if } \tau > 0 \end{cases} \qquad (3)$$

in cui c_{bw} rappresenta ancora il grado di attivazione, mentre c_{fw} può essere interpretato come grado di rilascio del movimento articolatorio. Facciamo notare, come appare in Figura 2, che per diversi valori del coefficiente c si ottiene una variazione dell'estensione della dominanza. Questa è una diretta conseguenza della variazione dell'andamento qualitativo delle curve, che corrisponde a differenti gradi di concavità o convessità. Una variazione del coefficiente θ, invece, è unicamente legata alla modifica dell'estensione del gesto articolatorio e non comporta variazioni delle proprietà qualitative delle curve della dominanza.



(a) (b)

Figura 2. Andamento della funzione di dominanza per diversi valori dei coefficienti θ_{fw} (con $c = 1$) (a) e c_{fw} (con $\theta = 15$) (b).

4.2 La funzione coarticolatoria

L'utilizzo di valori di attivazione e di rilascio differenti ci permette di dare rilievo ad alcune importanti proprietà del modello. Se analizziamo le caratteristiche di base della funzione coarticolatoria attraverso un modello semplificato a due target, considerando l'equazione (2), possiamo fare le seguenti osservazioni:

- Il momento in cui le dominanze di fonemi successivi si incrociano nel tempo corrisponde al raggiungimento di metà della distanza spaziale tra i due target corrispondenti. Infatti, definendo $D_0 = D_1(t_0) = D_2(t_0)$ il valore assunto dalle due dominanze in tale istante, possiamo scrivere in base alla (2):

$$F_0 = \frac{T_1 D_0 + T_2 D_0}{2 D_0} = \frac{T_1 + T_2}{2} \qquad (4)$$

In questo caso non viene considerato l'eventuale contributo di altri fonemi al di fuori di quelli considerati che potrebbero modificare questa situazione.

- Nella situazione in cui in t_o i valori delle dominanze sono diversi, rispettivamente $D_1 = D_1(t_0)$ e $D_2 = D_2(t_0)$ per il primo e secondo target, il valore finale della funzione coarticolatoria sarà dato da

$$F_0 = \frac{T_1 D_1 + T_2 D_2}{D_1 + D_2} \qquad (5)$$

- Consideriamo il caso $T_1 > T_2$; possiamo allora scrivere:

$$F_0 = T_2 + \frac{D_1}{D_1 + D_2}(T_1 - T_2) \qquad (6)$$

Da ciò possiamo dedurre che, fissati i target, il valore finale rimane invariato a patto che il rapporto $D_1/(D_1 + D_2)$ rimanga costante.

- A parità di differenza tra le dominanze $D_1 - D_2$, più piccolo sarà il valore di D_1, maggiormente ci avviciniamo al primo target (se $D_1 > D_2$) o al secondo target (se $D_2 > D_1$).

Quanto appena detto appare in Figura 3 in cui viene descritto il diverso comportamento della funzione coarticolatoria variando distintamente i coefficienti θ e c a partire dai valori base $\theta = 15$ e $c = 1$. In essa si è cercato di mantenere pressoché uguale il punto di incontro delle dominanze per coppie di curve ($\theta = 5$ a sinistra con $c = 1.7$ a destra; $\theta = 40$ a sinistra con $c = 0.7$ a destra). La differenza sostanziale degli andamenti nelle curve può essere osservata a partire dall'istante dell'incrocio delle dominanze in poi: un cambiamento del coefficiente θ ha maggiori effetti in prossimità del secondo target di quanto non si abbia cambiando il coefficiente c. In particolare nel caso di $\theta = 5$ l'influenza del primo target si mantiene forte sul secondo, mentre per $\theta = 40$ essa decade rapidamente. Con la variazione del coefficiente c si ha invece un comportamento più regolare.

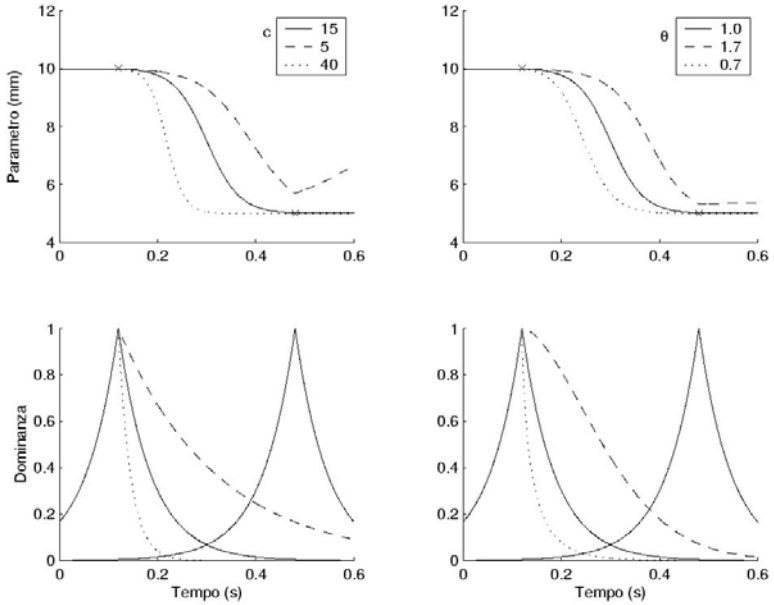


Figura 3. Variazione dei coefficienti θ_{fw} e c_{fw} e conseguente effetto sulla funzione di coarticolazione.

Distingueremo ora due casi principali di funzioni di dominanza, l'una di tipo concavo con coefficiente c pari a 1, l'altra di tipo convesso con coefficiente c pari a 2. Vediamo in Figura 4 cosa accade in caso di variazione della posizione temporale tra funzioni di dominanza. Il fatto di avere valori «alti» dell'incrocio delle dominanze comporta una transizione più diretta tra i target. Ciò vale fino a che l'estensione delle dominanza non va oltre il massimo dei target adiacenti. In questo caso l'andamento finale in prossimità di un target risente dell'influenza di quello adiacente ed il valore del target non viene raggiunto. L'utilizzo di funzioni di dominanza convesse fa sì che la situazione di incrocio «alto» venga raggiunta in più breve tempo. Tuttavia ciò vale anche per la situazione di perdita del target a causa della reciproca influenza delle dominanze.

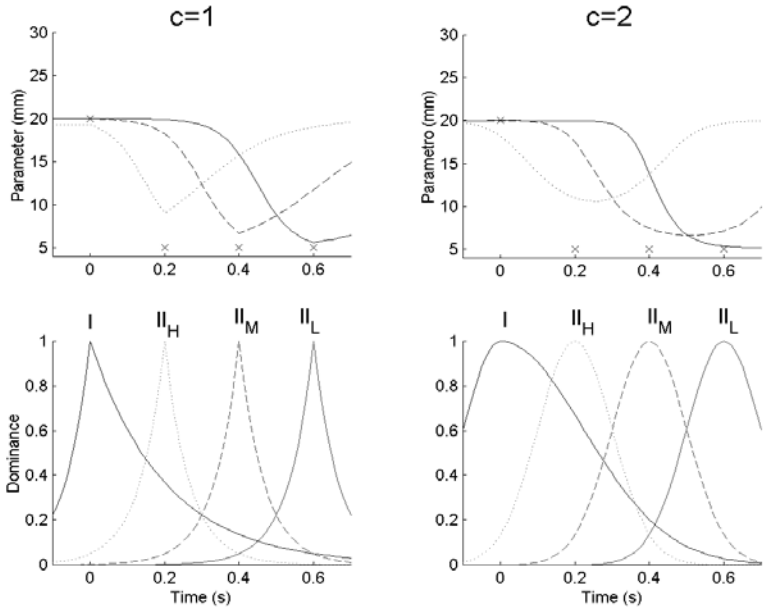


Figura 4. Andamento della funzione del modello risultante dalla sovrapposizione di due diverse tipologie di funzione di dominanza (c=1 e c=2) relative a due fonemi caratterizzati da differenti velocità di eloquio (II_H = elevata, II_M= media, II_L= bassa). I target articolatori sono posti rispettivamente a $20mm$ e $5mm$ per il primo e secondo target.

In figura 5 è rappresentato un modello a tre target che illustra cosa accade quando facciamo variare l'ampiezza della dominanza. L'influenza del segmento centrale decresce man mano che l'ampiezza della sua funzione di dominanza si abbassa. In questo caso l'uso di funzioni di dominanza convesse rende più regolare il comportamento della funzione finale nella zona centrale.

La caratteristica interessante del modello è che descrive bene le transizioni fisiche per quanto riguarda la sovrapposizione di gesti in presenza di consonanti neutre. In questo contesto, come scrivono gli stessi Cohen e Massaro, esso condivide l'idea di resistenza coarticolatoria definita da Bladon ed Al-Bamerni (Bladon & Al-Bamerni, 1976). Altrettanto non si può dire in presenza di gesti con alta resistenza che entrano in contrasto. In particolare con le consonanti bilabiali (/p b m/) e labiodentali (/f v/) un grosso problema è dato dalla perdita del target articolatorio ad elevate velocità di eloquio ("*speech rate*"), in modo analogo a quanto descritto in Figura 4.

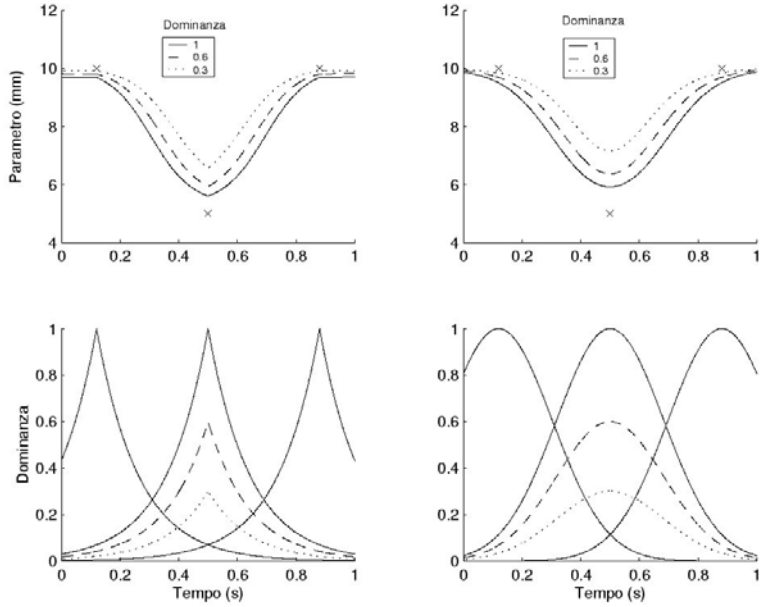


Figure 5. Variazione del coefficiente di dominanza α e risultante funzione coarticolatoria.

4.3 Funzione articolatoria e approssimazione locale

Le considerazioni che seguono prendono spunto dagli studi relativi ai modelli locali per l'approssimazione di funzioni, ed in particolare alle ricerche effettuate da Cleveland e Loader (Cleveland & Loader, 1995) e da Atkenson et alii (Atkenson et alii, 1985) nell'ambito della regressione locale pesata.

Creare un modello locale significa a livello basilare:

- separare lo spazio di origine dei dati in regioni;
- associare ad ogni regione una caratteristica che approssimi l'andamento desiderato.

A livello elementare possiamo pensare che la caratteristica sia costante e che le regioni siano contigue. Avremo allora un andamento a gradini del tipo mostrato in Figura 6. Immaginiamo, nella nostra situazione, che le regioni rappresentino i segmenti fonetici e che la caratteristica corrisponda al target articolatorio.

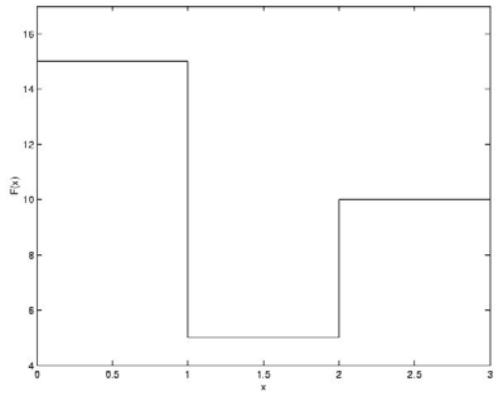


Figure 6. Approssimazione con andamento locale costante e regioni contigue

Per ovviare alle discontinuità tra le caratteristiche, estendiamo e facciamo sovrapporre le regioni ed inseriamo una funzione peso $W(x)$ che determina la transizione tra una caratteristica ed un'altra. Tale funzione avrà il massimo valore nella

zona in cui la caratteristica è dominante, mentre decadrà rapidamente a zero nelle zone di transizione adiacenti. Definiamo quindi l'attivazione della caratteristica i-esima centrata in x_i come il peso $W_i(x)$ normalizzato dalla somma dei pesi delle altre caratteristiche. Poniamo, poi, che l'estensione delle regioni e delle sovrapposizioni possa cambiare per le varie caratteristiche. L'attivazione, se il numero di regioni è pari a N, sarà quindi

$$A(x)_i = \frac{W_i\left(\frac{x-x_i}{h_i}\right)}{\sum_{i=1}^{N} W_i\left(\frac{x-x_i}{h_i}\right)} \qquad (7)$$

in cui h_i determina l'estensione della regione i-esima.

La funzione finale sarà la somma delle varie attivazioni moltiplicate per la caratteristica costante P_i

$$F(x) = \sum_{i=1}^{N} P_i A(x)_i = \frac{\sum_{i=1}^{N} P_i W_i\left(\frac{x-x_i}{h_i}\right)}{\sum_{i=1}^{N} W_i\left(\frac{x-x_i}{h_i}\right)}. \qquad (8)$$

Se poniamo $P_i = T_i$, $h_i = \frac{1}{\theta_i}$ e $W_i(x) = D_i(t)$ riotteniamo una funzione simile a quella descritta da Cohen e Massaro (vedi formula 2) e, se consideriamo una funzione peso di tipo gaussiano, il comportamento della (8) è del tutto analogo a quello descritto nel paragrafo precedente. La funzione peso $W(x)$ viene definita comunemente con il nome di Kernel. Nell'ambito della regressione locale, stiamo infatti parlando di Kernel Regression e la (8) prende il nome di stimatore di Nadaraya-Watson.

Un'approssimazione locale costante può tuttavia essere debole. Un caso più generale può essere ottenuto considerando una caratteristica locale che possa variare, ad esempio, all'interno di una famiglia di funzioni parametriche $P(x)$. Possiamo allora scrivere

$$F(x) = \frac{\sum_{i=1}^{N} P_i(x) W_i\left(\frac{x-x_i}{h_i}\right)}{\sum_{i=1}^{N} W_i\left(\frac{x-x_i}{h_i}\right)} \qquad (9)$$

Solitamente si utilizza la famiglia dei polinomi di secondo grado per cui

$$P_i(x) = a_2(x-x_i)^2 + a_1(x-x_i) + a_o. \qquad (10)$$

Questa famiglia di polinomi è sufficiente per ottenere delle buone approssimazioni, ma si può pensare al caso generale di una famiglia di polinomi di grado p. Possiamo vedere un esempio dell'effetto di tale variazione in Figura 7 dove si è utilizzata una funzione del tipo

$$Y(x) = a_1|x| + 1. \qquad (11)$$

Cleveland e Loader fanno notare (Cleveland & Loader, 1995) che un cambiamento, ad esempio, da un grado 1 ad un grado 2 può significare un cambiamento sostanziale dei risultati, per cui consigliano di considerare la famiglia dei polinomi misti, nei quali il grado non è più un numero naturale, ma un numero reale positivo. Definiscono polinomio misto di grado reale positivo non intero $p = m + c$, dove m è un intero e $0 < c < 1$, la media pesata dei polinomi di grado m e $m+1$, con peso $1-c$ per il primo e c per il secondo.

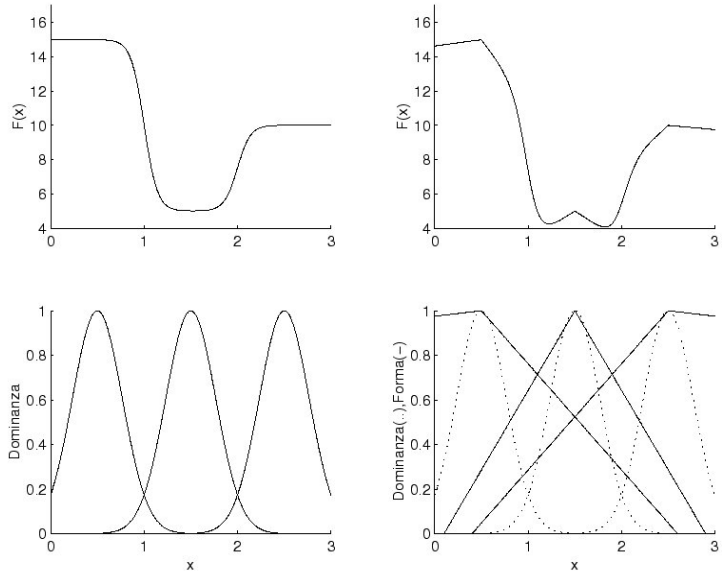


Figure 7. Influenza del polinomio sull'andamento della curva. La prima curva (in alto a sinistra) è generata facendo uso unicamente della funzione peso (in basso a sinistra), mentre la seconda (in alto a destra) è il risultato dell'uso combinato della funzione peso e della funzione 11 (in basso a destra).

4.4 Funzione di forma

In base a quanto illustrato nel paragrafo precedente, per migliorare la precisione dell'approssimazione, integriamo nel modello coarticolatorio una funzione che riprende il concetto espresso da Cleveland e Loader (Cleveland & Loader, 1995) . Vogliamo che tale funzione esprima il concetto di polinomio di grado reale e che allo stesso tempo sia il più semplice possibile. L'espressione che utilizzeremo sarà quindi del tipo

$$Y(x) = a|x|^p + 1 \qquad (12)$$

con p reale positivo.

Per il nostro modello in particolare utilizzeremo una forma che caratterizza in modo diverso l'andamento del polinomio secondo la direzione di estensione della coarticolazione:

$$S_{TL}(\tau) = \begin{cases} \beta_{bw}|\tau|^{p_{bw}} + 1 \text{ se } \tau < 0 \\ \beta_{fw}|\tau|^{p_{fw}} + 1 \text{ se } \tau > 0 \end{cases} \qquad (13)$$

dove TL sta per «Time-Locked», in quanto la sua caratteristica è indipendente dal posizionamento dei fonemi precedenti e successivi, analogamente a ciò che accade nel modello di Bell-Berti e Harris (Bell-Berti & Harris, 1981).

Chiameremo l'espressione definita in (13) *funzione di forma*, perché il suo effetto principale è quello di modellare l'andamento del target articolatorio in prossimità del suo massimo rilievo. Avremo quindi un target articolatorio non più discreto, ma che si evolve nel tempo con una propria caratteristica.

Si può anche pensare ad una forma di tipo «Look-Ahead» in cui l'influenza della funzione di forma sia proporzionale alla distanza con il target successivo o antecedente. Possiamo allora scrivere

$$S_{LA}(\tau) = \begin{cases} \beta_{bw} \left| \dfrac{\tau}{h_{bw}} \right|^{p_{bw}} + 1 \text{ se } \tau < 0 \\ \beta_{fw} \left| \dfrac{\tau}{h_{fw}} \right|^{p_{fw}} + 1 \text{ se } \tau > 0 \end{cases} \quad (14)$$

in cui h_{bw} e h_{fw} rappresentano fattori proporzionali alla distanza dai target precedenti o successivi.

L'utilizzo delle funzione di forma risulta utile nel riprodurre andamenti con caratteristiche particolari come ad esempio la pendenza rilevata nella produzione della vocale /u/ in alcuni contesti consonantici come illustrato ad esempio in Figura 8.

Tuttavia il ruolo fondamentale viene svolto in situazioni in cui si necessita di un rapido smorzamento come, ad esempio, nel caso del rilascio del gesto alla fine di una frase, come illustrato in Figura 9. In questo caso particolare senza l'utilizzo di tale funzione non sarebbe possibile in alcun modo riprodurne l'andamento.

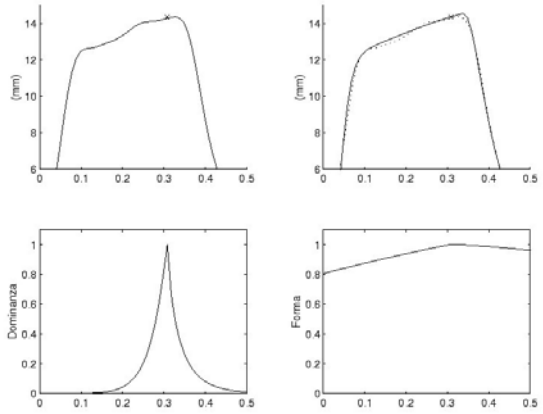


Figure 8. Esempio di andamento del labbro inferiore nella produzione della vocale /'u/ (in alto a sinistra) e relativa descrizione attraverso l'integrazione della funzione di forma nel modello (in alto a destra, in tratteggio abbiamo il raffronto con l'andamento reale). L'andamento della dominanza e della funzione forma sono in basso (rispettivamente a sinistra e destra).

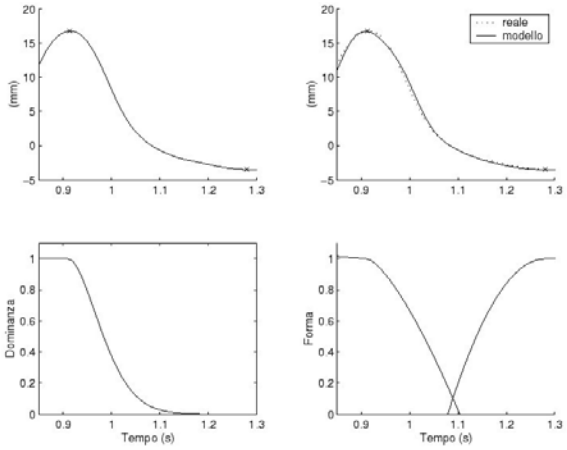


Figure 9. Esempio di rilascio del gesto vocalico (in alto a destra) e descrizione tramite il modello coarticolatorio (in alto a sinistra). Sotto sono rappresentate le dominanze e la funzione forma utilizzate; è stata rappresentata anche la funzione di forma del secondo target perché gioca un ruolo importante per l'andamento finale.

4.4 Funzione di resistenza temporale

Per ovviare al problema della perdita del target consonantico si estende temporalmente il concetto di resistenza coarticolatoria, precedentemente introdotto, tramite l'introduzione di una variazione dell'ampiezza della funzione di dominanza. Si vuole in tal modo avere la possibilità di bloccare i gesti articolatori, relativi sia al fonema precedente che a quello successivo, in modo da annullarne la reciproca influenza e di conseguenza imporre il raggiungimento forzato del target. Per far questo, ad ogni dominanza è stato associato un esponenziale negativo $R(\tau)$, denominato funzione di "*resistenza temporale*", con un'andamento simile alla dominanza, ma con estensione variabile in base alla collocazione dei fonemi precedenti o successivi ed al loro grado di resistenza. Dopo alcune prove su dati reali abbiamo scelto una funzione con andamento esponenziale del tipo

$$R(\tau) = \begin{cases} e^{-6\left|\frac{\tau}{h_{bw}}\right|^4} & \text{se } \tau < 0 \\ e^{-6\left|\frac{\tau}{h_{fw}}\right|^4} & \text{se } \tau > 0 \end{cases} \qquad (15)$$

in cui h_{bw} e h_{fw} hanno un significato analogo a quelli della funzione forma (si veda Figura 10).

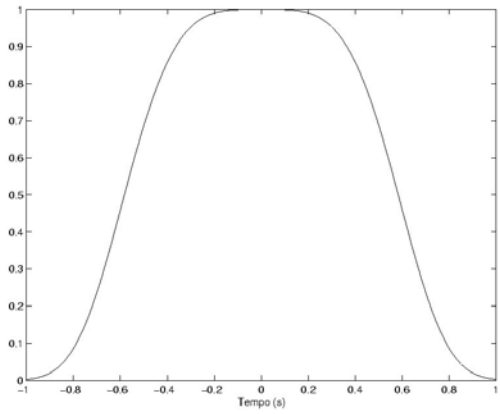


Figura 10. Andamento della funzione di resistenza temporale per $h_{bw}, h_{fw} = 1$.

Ad ogni fonema è associato un coefficiente di resistenza k_r attraverso il quale è possibile calcolare ricorsivamente il valore di h_{bw} e h_{fw}. Se consideriamo, infatti, il fonema i-*esimo* con target articolatorio al tempo t_i, il valore di $h_{\mathrm{fw_i}}$ si ottiene mediante la seguente procedura ricorsiva:

[Passo 1]

se il fonema (i+1)*-esimo* è caratterizzato da $k_{\mathrm{R}} = 1$,

$$h_{\mathrm{fw_i}} = (t_{i+1} - t_i), \text{ altrimenti salta a [Passo 2];} \qquad (16)$$

[Passo 2]

$$h_{\mathrm{fw_i}} = (t_{i+1} - t_i) + k_R \cdot (h_{\mathrm{fw_{i+1}}}).$$

In pratica [Passo1] se il coefficiente di resistenza k_r del segmento successivo è pari a 1, allora h_{fw} è pari alla distanza tra il target attuale e quello successivo, altrimenti [Passo 2] h_{fw} è uguale alla distanza tra il target attuale e quello successivo sommato al valore di h_{fw} del segmento successivo calcolato con la medesima procedura.

La procedura per h_{bw_i} si ottiene sostituendo (i+1) con (i-1).

In altre parole, se la resistenza del fonema successivo è massima ($k_{\text{R}} = 1$), 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 (vedi Figura 11). Per $k_{\text{R}} < 1$, l'estensione di $R(\tau)$ aumenta seguendo la procedura ricorsiva sopra introdotta.

Si deve poi sottolineare che, per limitare il numero di ricorsioni, si pone il vincolo che k_r debba essere maggiore di un valore di poco inferiore ad 1. Per k_r = 0.97, valore ottenuto euristicamente, si assicura l'esecuzione in media di tre o quattro ricorsioni.

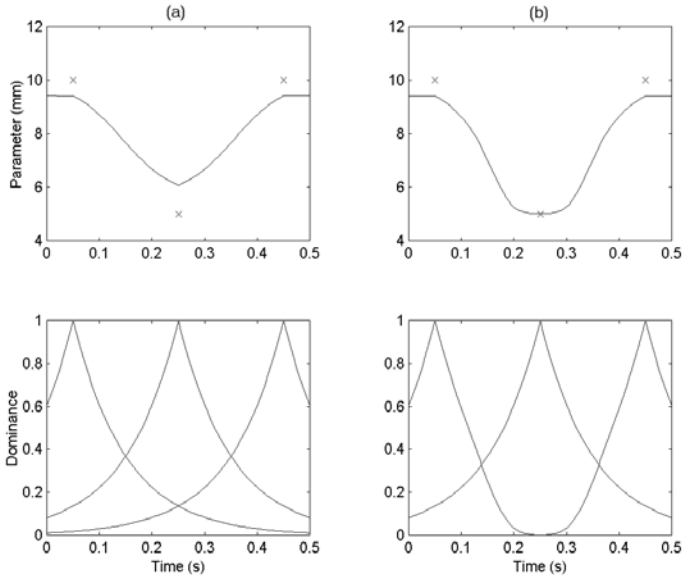


Figura 11. Esempio degli effetti coarticolatori riscontrabili in presenza di fonemi ad elevato valore di dominanza e pronunciati velocemente simulati dal modello di coarticolazione originario (a) e da quello modificato mediante l'introduzione della funzione di resistenza temporale (b). Si noti come, nel secondo caso il target centrale è perfettamente raggiunto.

4.4 Funzione coarticolatoria finale

Riassumendo infine tutte le modifiche apportate al modello originale la funzione coarticolatoria risultante è data da:

$$F_{new}(t) = \frac{\sum_{i=1}^{N} T_i \cdot S_{LA_i}(t - t_i) \cdot R_i(t - t_i) \cdot D_i(t - t_i)}{\sum_{i=1}^{N} R_i(t - t_i) \cdot D_i(t - t_i)} \qquad (17)$$

Il termine $S_{LA_i}(t-t_i)$ non compare a denominatore, analogamente a quanto accade nella (9), mentre essendo $R_i(t-t_i)$ in stretta relazione con la dominanza entrerà anch'esso a far parte del termine di normalizzazione. Si può notare, infine, che con l'introduzione della funzione di resistenza temporale, il ruolo svolto dal coefficienti di attivazione e rilascio diventa fondamentale. Si pensi, ad esempio, a due fonemi successivi entrambi con resistenza unitaria. Solo una variazione consistente del coefficiente θ comporta variazioni significative nell'andamento finale. Ciò non vale invece per i valori di c_{fw} e c_{bw}. Si può infine intuire che è agendo principalmente su di essi che riusciamo a diversificare le caratteristiche cinematiche in presenza di fonemi con alta resistenza.

5. STIMA DEI PARAMETRI DEL MODELLO

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 una serie di 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 (Ferrigno & Pedotti, 1985). La procedura di stima dei parametri è basata su metodo classico di minimizzazione dei minimi quadrati:

$$e(x)=\sum_{n=1}^{N}\left[Y(n)-F(n,x)\right]^2 \qquad (18)$$

fra i dati reali (*Y(n)*) e le curve ottenute in uscita dal modello modificato (*F(n,x)*), rappresentato dalla funzione illustrata in (17), per 5 ripetizioni dello stesso tipo di sequenze. In (18), N rappresenta il numero totale di ripetizioni, *F(n,x)* il valore dell'n-esimo campione e x il vettore dei coefficienti per l'intero insieme di fonemi coinvolto. Nel caso di sequenze VCV, il vettore x è costituito da tre sequenze concatenate di coefficienti organizzati in matrici simili a quella illustrata in tabella 2.

N.	*Coefficient*	/p/	/b/	/m/
1	t	0.330	0.260	0.140
2	T	−2.000	−1.000	−0.200
3	α	1.000	1.000	1.000
4	k	1.000	1.000	1.000
5	θ_{bw}	8.907	8.907	8.907
6	θ_{fw}	3.000	3.000	3.000
7	c_{bw}	0.881	0.584	0.386
8	c_{fw}	2.072	2.072	0.595
9	p_{bw}	1.000	1.000	1.000
10	p_{fw}	1.000	1.000	1.000
11	β_{bw}	0.025	0.025	0.025
12	β_{fw}	0.025	0.025	0.025

Tabella2. Organizzazione dei coefficienti del modello per l'algoritmo di ottimizzazione.

Il calcolo dei coefficienti viene eseguito in passi successivi che combinano analisi manuali e tecniche automatiche di ottimizzazione. Non è infatti possibile trattare i dati in modo completamente automatico, in quanto per il modello utilizzato, a differenza ad

esempio di quanto accade nel lavoro di Le Goff (Le Goff, 1997), la funzione costo presenta molti minimi globali e deve essere necessariamente guidata in modo manuale verso gli opportuni valori target finali. Si sono riscontrati rari casi di minimi locali indesiderati a cui però si è potuto porre rimedio mediante una correzione dei valori iniziali di partenza del processo di minimizzazione. Particolare attenzione è stata rivolta alla selezione del metodo di ottimizzazione. Essendo il numero di parametri in gioco molto alto, si è presentata la necessità di sviluppare un algoritmo che avesse la proprietà di convergere velocemente verso il minimo in poche iterazioni. E' stato scelto un metodo di tipo *Trust Region* con approssimazione del passo di aggiornamento in un sottospazio a due dimensioni che contenga il cammino calcolato secondo il metodo Dogleg (Schultz et alii, 1985). Tale metodo ha infatti una forte convergenza e garantisce, nel nostro caso, una buona approssimazione del minimo già in 10-15 iterazioni.

Nella Figura 12 è illustrato il risultato dell'applicazione del modello in un semplice esempio relativo all'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 globale fra le traiettorie reali e quelle simulate risulta essere inferiore a 0.3 mm.

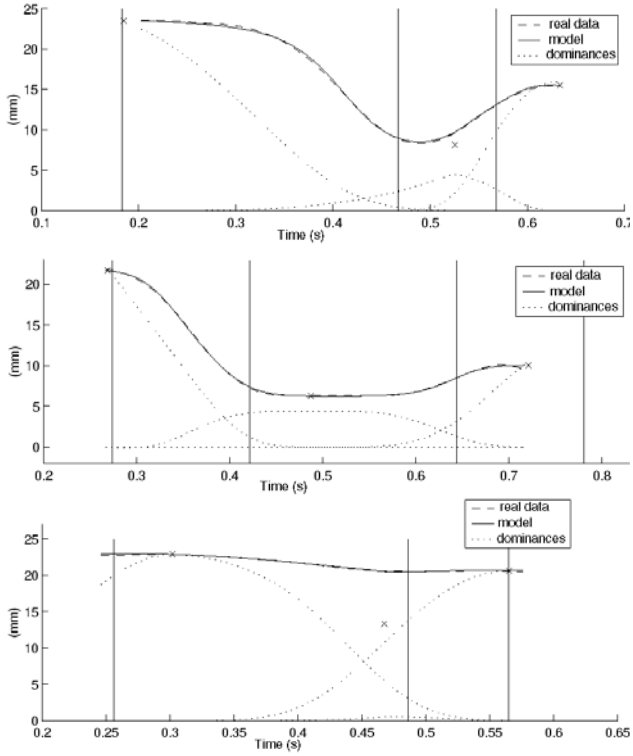


Figura 12. Traiettorie del parametro di apertura del labbro inferiore durante la produzione delle sequenze isolate /'a d a/ (a), /'a dz a/ (b) e /'a l a/ (c). Le linee tratteggiate indicano le funzioni di dominanza.

6. LE APPLICAZIONI: GRETA E LUCIA

Il modello è stato applicato con successo a GRETA (Pasquariello, 2000; Pelachaud et alii, 2001) e più recentemente a LUCIA (Cosi, 2002, Cosi, 2003), due facce parlanti in italiano emotive ed espressive, il cui diagramma a blocchi è illustrato in Figura 13..

GRETA e LUCIA, che parlano mediante la versione italiana di FESTIVAL (Cosi et alii, 2001), sono dei motori di animazione facciale entrambe compatibili con lo standard

MPEG-4 (MPEG4, 2003) ed utilizzabili per realizzare un *decoder* compatibile con il cosiddetto "*Predictable Facial Animation Object Profile*". 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.

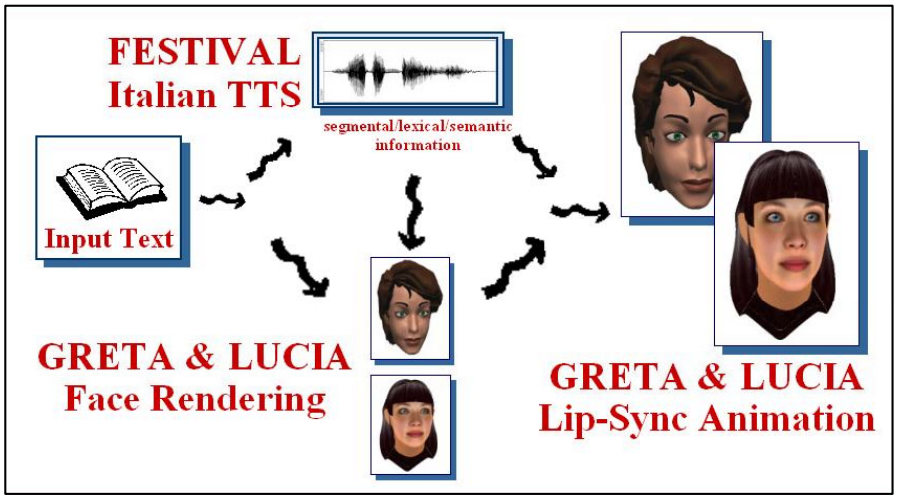


Figura 13. Diagramma a blocchi di Greta and Lucia, due facce parlanti in italiano.

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.

La differenza principale fra i due modelli risiede essenzialmente nell'utilizzo di *texture* reali per LUCIA (vedi Figura 14), 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 (VRML97, 1997). GRETA e LUCIA sono entrambe 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.

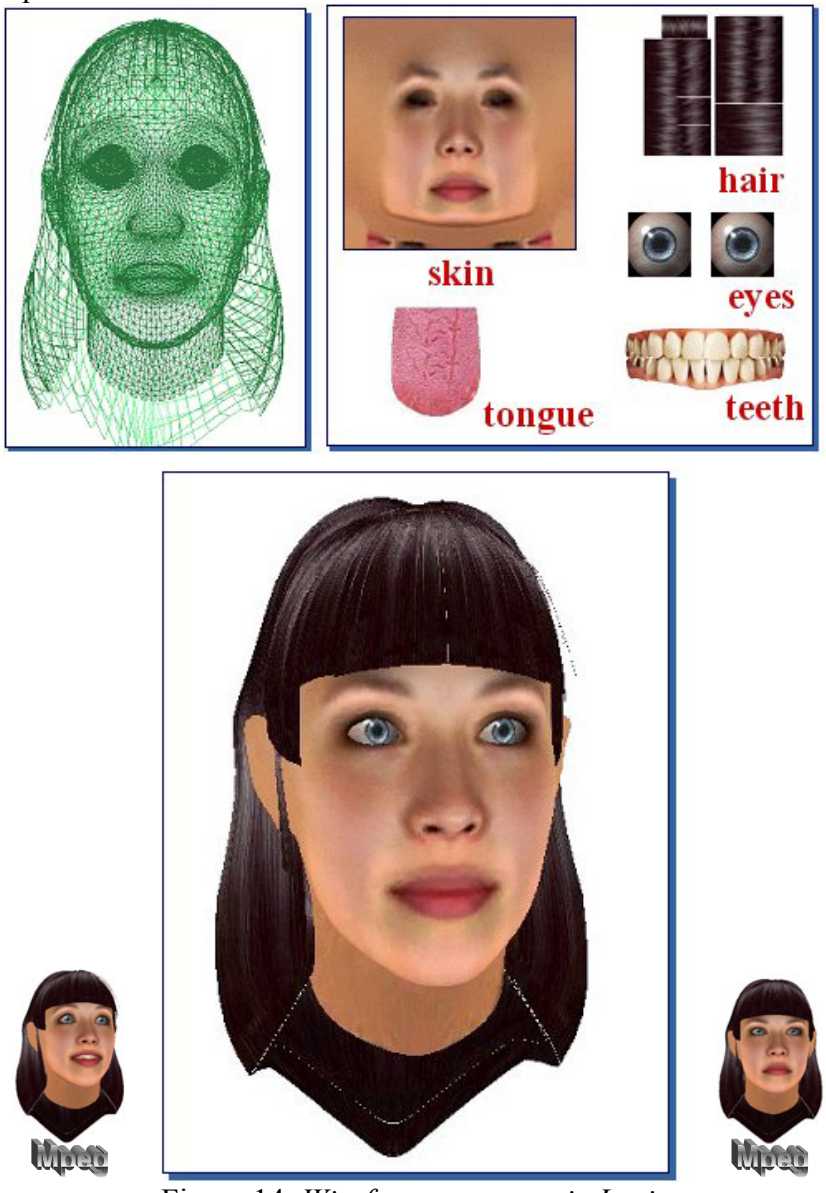


Figura 14. *Wireframe e texture in Lucia.*

(per visualizzare due esempi di animazione "*cliccare*" con il mouse sulle due faccine laterali)

7. 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.

BIBLIOGRAFIA

Abry C., Lallouache M.T. (1991), Does increasing representational complexity lead to more speech variability?, *Phonetic Experimental Research at the Institute of Linguistics*, University of Stockholm, 1991, number 14, pages 1-5.

Al-Bamerni A., Blandon A. (1991), One-stage and two-stage patterns of velar coarticulation, *Journal of the Acoustical Society of America*, 1982, Vol. 72, number 104.

Atkenson C.G., Moore A.W., S. Schaal (1997), Locally weighted learning, *Artificial Intelligence Review*, 1997, Vol. 11, pages 11-73.

Bell-Berti F., Harris K.S. (1981), A temporal model of speech production, *Phonetica*, 1981, Vol. 38, pages 9-20.

Beskow J. (1995), Rule-Based Visual Speech Synthesis, in *Proceedings of Eurospeech '95, 4th European Conference on Speech Communication and Technology*, Madrid, September 1995.

Bladon, R.A., Al-Bamerni, A. (1976), Coarticulation Resistance in English \l\, *Journal of Phonetics*, 4, 1976, pages. 135-150.

Bregler C., Covell M., Slaney M. (1997), Video Rewrite: Driving Visual Speech with Audio, in *Proceedings of SIGGRAPH '97*, 1997, pages 353-360.

Cleveland W.S., Loader C. (1996), Smoothing by local regression: principles and methods, in *Statistical theory and computational aspects of smoothing* (Ed.. W. Hardel and M. Schimek), Physica-Verlag, 1996, pages. 10-49.

Chomsky N., Halle M. (1968), *The Sound Pattern of English*, Harper and Row, New York, NY, 1968.

Cohen, M. M., Beskow, J., Massaro, D.W. (1998), Recent Developments in Facial Animation: An Inside View, in *Proceedings of the International Conference on Auditory-Visual Speech Processing - AVSP'98*, December 4-6,1998, Terrigal, Australia, pages 201-206.

Cohen M., Massaro D. (1993), Modeling Coarticulation in Synthetic Visual Speech, in *Models and Techniques in Computer Animation*, Magnenat-Thalmann N., Thalmann D. (Editors), Springer Verlag, Tokyo, 1993, pages 139-156.

Cosi P., Tesser F., Gretter R., Avesani C. (2001), Festival Speaks Italian!, in *Proceedings of Eurospeech 2001*, Aalborg, Denmark, September 3-7 2001, pp. 509-512.

Cosi P., Ferrari V., Magno Caldognetto E., Perin G., Tisato G., Zmarich C. (2002), GRETA e LUCIA: due realistiche facce parlanti animate mediante un nuovo modello di coarticolazione, in *Atti delle XIII Giornate di Studio GFS 2002*, Pisa, 28-30 November, 2002 (in press).

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*).

Daniloff R., Moll K. (1973), On defining coarticulation, *Journal of Speech and Hearing Research*, 1973, Vol. 1, pages 239-248.

Farnetani E., Recasens D. (1999), Coarticulation Models in Recent Speech Production Theoories, in *Coarticulation in Speech Production*, Hardcastle W.J. (Editors), Cambridge University Press, Cambridge, 1999.

Ferrigno G., Pedotti A. (1985), ELITE: A Digital Dedicated Hardware System for Movement Analysis via Real-Time TV Signal Processing, in *IEEE Transactions on Biomedical Engineering*, BME-32, 1985, pages 943-950.

Henke W.L. (1966), Dynamic articulatory model of speech production using computer simulation, *Unpublished doctoral dissertation*, MIT Cambridge, Ma, 1966.

Keating P.A. (1988), The window model of coarticulation: articulatory evidence. *UCLA Working Papers in Phonetics*, 69:3–29, 1988.

Keating P.A. (1990), The window model of coarticulation: articulatory evidence. In M.E. Beckam, (eds.), *Papers in Laboratory Phonetics I: between the grammar and the physics of speech*, pages 451–470. Cambridge University Press, 1990.

Kozhevnikov V., Chistovich L. (1965), Speech: Articulation and perception, *Joint Publications Research Service*, Washington, DC, 1965, Vol. 30, series 534.

Lee Y., Terzopoulos D., Waters K. (1995), Realistic Face Modeling for Animation, in *Proceedings of SIGGRAPH '95*, 1995, pages 55–62.

Le Goff *B. (1997), Synthèse à partir du texte de visages 3D parlant français*, PhD thesis, Grenoble, France, October 1997.

Le Goff B. and Benoit C. (1996), A text-to-audiovisual speech synthesizer for French. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP '96)*, Philadelphia, USA.

Lindblom B. (1963), "On Vowel Reduction", *Technical Report 29, KTH, The Royal Institute of Technology*, Speech Transmission Laboratory, Stockholm, 1963.

Löfqvist, A. (1990), Speech as Audible Gestures, in *Speech Production and Speech Modeling*, Hardcastle W.J., Marchal A. (Editors.), Dordrecht: Kluwer Academic Publishers, 1990, pages. 289-322.

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

MPEG-4 standard (2003), http://mpeg.telecomitalialab.com/standards/mpeg4.

Munhall K.G., Löfqvist A. (1992), Gestural aggregation in speech: laryngeal gestures, *Journal of Phonetics*, 1992, Vol. 20, pages 111-126.

Öhman S. (1966), "Coarticulation in VCV utterances: spectrographic measurements", *Journal of acoustical Society of America*, 1966, Vol. 39, pages 151-168.

Öhman S. (1967), "Numerical model of coarticulation", *Journal of Acoustical Society of America*, 1967, Vol. 41, pages 310-320.

Pasquariello, S. (2000), *Modello per l'animazione facciale in MPEG-4*, M.S. thesis, University of Rome, 2000.

Pelachaud C., Magno Caldognetto E., Zmarich C., Cosi P. (2001), Modelling an Italian Talking Head, in *Proceedings of AVSP 2001*, Aalborg, Denmark, Settembre 7-9 2001, pages 72-77.

Salzman E.L., Munhall K.G. (1989), A dynamical approach to gestural patterning in speech production, *Ecological Psychology*, 1989, Vol. 1, number 4, pages 333-382.

Schultz G.A., Schnabel R.S., Byrd R.H. (1985), A family of trust-region-based algorithms for unconstrained optimization with strong global convergence properties, *SIAM Journal on Numerical Analysis}*, 1985, Volume 22, pages 47-67.

Vatikiotis-Bateson E., Munhall K.G., Hirayama M., Kasahara Y., Yehia H. (1996), Physiology-Based Synthesis of Audiovisual Speech, in *Proceedings of 4th Speech Production Seminar: Models and Data*, 1996, pages 241-244.

VRML97 (1997), http://www.web3d.org/Specifications/VRML97/