# NUOVI STRUMENTI DI *INTERFACE* PER L'ELABORAZIONE DELLE FACCE PARLANTI

Graziano Tisato, Piero Cosi, Giacomo Sommavilla, Claudio Zmarich
ISTC – SPFD – CNR
Istituto di Scienze e Tecnologie della Cognizione
Sezione di Padova - Fonetica e Dialettologia
Via Martiri della Libertà 2 - 35137 Padova
{tisato, cosi, sommavilla, zmarich}@ pd.istc.cnr.it

## 1. SOMMARIO

Questo articolo presenta gli sviluppi più recenti di **InterFace**, un *software* interattivo realizzato in Matlab all'ISTC-SPFD, per l'animazione audio-visuale delle **Facce Parlanti**. Per completezza di informazione, questo testo riprende ed integra in maniera esaustiva le presentazioni parziali già fatte precedentemente (Tisato *et alii*, 2005a, Tisato *et alii*, 2005b, Cosi *et alii*, 2005).

La ricerca nel campo delle teorie di produzione e percezione della lingua parlata, del riconoscimento della voce, degli agenti conversazionali, dell'insegnamento delle lingue, della riabilitazione della voce, dello studio delle emozioni, ecc., deve far fronte a necessità sempre crescenti di elaborazione di dati articolatori ed acustici.

La realizzazione di InterFace intende rispondere a queste esigenze con lo sviluppo di strumenti *software* adeguati alla complessità delle problematiche implicate.

Per quanto riguarda in particolare i dati articolatori, InterFace permette di:

- Estrarre le traiettorie 3D provenienti da sistemi di *Motion Capture*, e sottoporle a elaborazione come: filtraggio del rumore, eliminazione delle componenti dovute alla rototraslazione, riscalamento alle dimensioni della faccia da animare.
- Definire un insieme di misure sulle traiettorie per ottenere i parametri articolatori voluti (ad es. apertura labiale, arrotondamento, protrusione, aggrottamento, asimmetrie labiali, ecc., con le relative misure di velocità ed accelerazione).
- Ricavare da quelle stesse traiettorie articolatorie una modellizzazione dei parametri rilevanti dal punto di vista linguistico, che tenga in debito conto i fenomeni di coarticolazione.
- Generare il flusso dei dati audio-visuali necessari all'animazione di un agente conversazionale, capace di esprimere emozioni.

Per quanto riguarda, d'altra parte, le Facce Parlanti, il sistema può arrivare ad un *set* di parametri di sintesi, partendo da quattro differenti tipi di dati (Fig. 1):

- **Dati reali** provenienti da sistemi di cattura degli andamenti cinematici dell'articolazione facciale. L'elaborazione di questi dati permette di realizzare una tipica ***Data-Driven Synthesis***.

- **Dati testuali** (puro testo o testo XML), da cui generare il flusso di dati audio-video di controllo dell'animazione facciale. Seguendo questo via, si ottiene una ***Text-to-Animation Synthesis***, o anche, nel caso dell'XML, una ***Symbolic-Driven Synthesis***.
- **Dati audio** elaborati in modo da ricavare la segmentazione fonetica del parlato con un sistema di riconoscimento automatico, e ottenere così la sincronizzazione dell'animazione con un audio preesistente. Questa modalità può essere chiamata una ***Wav-to-Animation Synthesis***.
- **Dati a basso livello**, per controllare manualmente il movimento di uno o più parametri di animazione e verificarne l'effetto con la sintesi video. Quest'ultimo procedimento si può definire come una ***Manual-Driven Synthesis***.

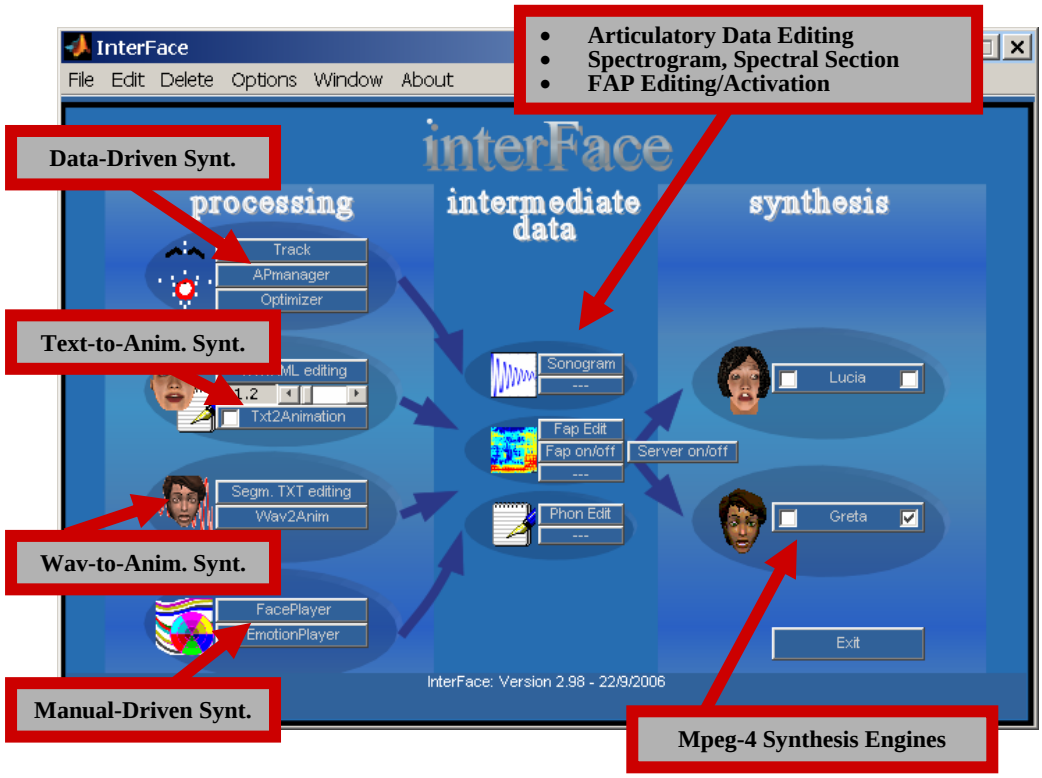


Figura 1 – Schermata principale di InterFace, che mostra una divisione orizzontale in tre aree funzionali: *Elaborazione*, *Editing dei dati audio-visuali* e *Sintesi*. Nella prima, si preparano i parametri per 4 diversi tipi di animazione.

## 2. INTRODUZIONE

Un paradigma, che trova attualmente largo consenso nella comunità scientifica, sostiene che la natura della comunicazione parlata e della trasmissione delle emozioni è inerentemente multimodale. Secondo questo principio, le informazioni linguistiche e paralinguistiche, provenienti tanto dal canale uditivo che da quello visivo si integrano a

vicenda, agevolando il processo comunicativo. Una prova viene dall'esperienza di tutti i giorni, per cui si verifica che l'intellegibilità del linguaggio aumenta notevolmente, anche in condizioni di ascolto pessime, quando si possa vedere la faccia dell'interlocutore che ci parla. Da un punto di vista linguistico, test di intellegibilità hanno portato all'individuazione di una serie di visemi, che sono l'equivalente visivo dei fonemi (si veda per l'italiano: Magno C. *et alii*, 1993, 1997, 1998, Cosi & Magno C., 2002).

Queste considerazioni spiegano come la ricerca si sia focalizzata da qualche decennio nel campo della trasmissione e della percezione audio-visuale, o bimodale, del linguaggio, sia dal punto di vista teorico, per la formulazione di teorie che ne spieghino i meccanismi di funzionamento (McGurk & MacDonald, 1976; Summerfield, 1987; Massaro, 1987,1996, 1998; Stork & Henneke, 1996), sia per fornire il supporto ai vari settori tecnologici, che si occupano dell'interazione uomo-macchina (Stork & Henneke, 1996; Massaro *et alii*, 2000; Magno C. *et alii*, 2001; Prendinger & Ishizuka, 2004), come, ad esempio, il riconoscimento automatico (Walden, 1977; Petajan, 1984; Stork *et alii*, 1992; Silsbee, 1993; Adjoudani & Benoit, 1995), la sintesi (Parke, 1974, 1982; Cohen & Massaro, 1990; Benoit *et alii*, 1992, 1996; Le Goff *et alii*, 1994, 1996; Le Goff, 1997; Lee *et alii*,1995; Beskow, 1995; Vatikiotis-Bateson, 1996; Massaro *et alii*, 2000; Pelachaud, 2001, Cosi *et alii*, 2001, 2002c, 2003a), le telecomunicazioni (Stork & Henneke, 1996; Chen & Rao, 1998); l'insegnamento (Cosi & Magno C., 2003b; Biscetti *et alii*, 2004) e la riabilitazione del linguaggio (Cosi *et alii*, 2004a).

A questo si è aggiunga negli ultimi anni una attenzione crescente per lo studio delle emozioni, che deriva dalla riconosciuta importanza del condizionamento, conscio e inconscio, che esse esercitano nelle relazioni interpersonali e nello sviluppo umano (Ekman & Friesen, 1977, 1978; Massaro & Egan, 1998; Cohen *et alii*, 1998; Cowie *et alii*, 2001; Douglas_Cowie & Campbell, 2003; Keltner, 2003).

Per quanto riguarda le applicazioni tecnologiche, è diffusa l'opinione che l'introduzione della bimodalità e delle emozioni possa migliorare in modo sostanziale l'efficacia e la naturalezza della comunicazione uomo-macchina. Mentre tradizionalmente i dati acustici e quelli visivi, tanto del parlato che delle emozioni, sono stati analizzati ed utilizzati separatamente (Cahn, 1990; Banse & Scherer, 1996; Scherer, 2003; Scherer *et alii*, 2003), le ricerche più recenti hanno introdotto la bimodalità e l'espressione delle emozioni negli agenti conversazionali (Pasquariello, 2000; Pelachaud, 2001; Bilvi, 2002; Magno C. *et alii*, 2003, 2004; Drioli *et alii*, 2003, 2004).

Per quanto riguarda l'audio, gli studi del passato hanno portato ad una categoria di sintetizzatori vocali (sistemi ad unità variabili, o *Automatic Unit Selection*) in grado di generare una voce decisamente più naturale e fluente dei sistemi precedenti (vedi ad esempio www.loquendo.it), che erano basati su la concatenazione di un numero limitato di difoni, oppure sul modello sorgente-filtro. Come i precedenti, comunque, anche i sistemi ad unità variabili risentono di un *handicap* implicito nella loro architettura, e cioè la mancanza di capacità espressive. Sia che si tratti della lettura del telegiornale o del racconto di una favola, che si tratti di una notizia buona o di una cattiva, il sintetizzatore in questione continuerà a produrre lo stesso identico flusso sonoro nelle diverse circostanze. I tentativi di sintesi audio più recenti vanno in due direzioni: la prima prevede l'acquisizione di *corpora* di dati per i sistemi ad unità variabili già predisposti per un certo numero di emozioni (Iida *et alii*,2003). La seconda segue la strada di aggiungere al motore di sintesi gli opportuni controlli sia per i parametri prosodici (Tesser *et alii*, 2003, 2004) che per quelli spettrali

(Rank & Pirker, 1998, Drioli *et alii*, 2003, 2004), che si adattino in maniera flessibile al significato del messaggio e allo stato emotivo che devono veicolare.

Le linee di tendenza suddette, e cioè bimodalità ed emozioni, hanno avuto due conseguenze immediate: da un lato, l'esigenza di acquisire grandi quantità di dati audio-visuali, dal momento che essi risultano essere specifici di ogni lingua e dialetto, e anche di ogni emozione (Cosi *et alii*, 2002b). Dall'altro, la necessità che ci sia sincronia fra dati acustici e visivi, dal momento che si vogliono comprendere il ruolo e l'influenza reciproca dei parametri nella produzione e percezione della comunicazione umana.

Una esigenza ulteriore è legata alla quantità dei dati da processare, che nel caso di un sistema a difoni tradizionale è normalmente dell'ordine delle migliaia di stimoli per ogni lingua, dialetto, o emozione presa in considerazione. Nel nostro caso, il *software*, messo a disposizione con il sistema ELITE (*ELaboratore di Immagini TElevisive*) (Ferrigno & Pedotti, 1985; www.bts.it), imponeva di configurare manualmente la maschera di riferimento dei punti facciali, di cui il programma doveva ricostruire il movimento, per ognuno delle migliaia di *file* acquisiti. A questo si doveva sommare una fase molto più lunga e faticosa di interventi manuali necessari per la correzione degli errori molto frequenti di tracciamento delle traiettorie, dovuti alla vicinanza dei punti e alla loro rapidità di movimento. Il tempo richiesto dalla elaborazione degli stimoli si traduceva in parecchi mesi di lavoro certosino, che metteva a dura prova la pazienza e gli occhi del malcapitato operatore.

Sono questi alcuni motivi che hanno portato alla creazione di un *software* apposito, chiamato InterFace (www.pd.istc.cnr.it/interface).

I vantaggi immediati di questo *software* si possono misurare nella riduzione dei tempi di elaborazione, che erano dell'ordine di molti mesi, e che sono passati ora a pochi giorni, e la drastica contrazione dei tempi di realizzazione e di test dell'animazione delle Facce Parlanti.

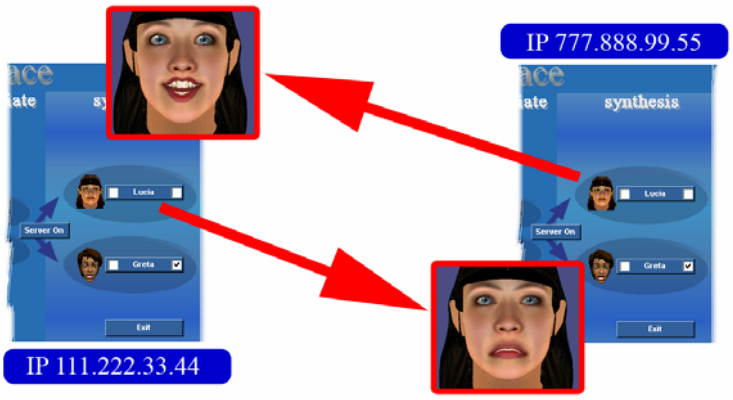


Figura 2 – Animazione di Talking Heads in modalità server.

## 3. INTERFACE

InterFace è un ambiente di sviluppo interattivo e flessibile, che è stato realizzato all'ISTC-SPFD con lo scopo di accelerare l'analisi dei dati bimodali, e l'estrazione, la modellizzazione ed il test dei parametri necessari all'animazione audio-visuale.

In Fig. 1 si può vedere la schermata principale del programma, in cui si nota la barra con i menu in alto, e la divisione in tre blocchi funzionali principali, con l'elaborazione dei dati a sinistra, la zona di *editing* e controllo dei dati intermedi al centro, e a destra la sintesi con gli agenti Greta e Lucia (www.pd.istc.cnr.it/lucia).

Nella barra dei menu, il pulsante *Options* permette di configurare le differenti applicazioni (compresa la lingua utilizzata) e di salvare i parametri in un *file* di inizializzazione.

L'area dell'elaborazione è la parte fondamentale di InterFace. Contiene le applicazioni sviluppate per trattare i dati, suddivisi secondo la loro tipologia, e cioè: **Dati bimodali** (vedi cap. 4-6), **Dati XML e testuali** (cap. 11), **Dati audio** (cap. 12), e **Dati a basso livello** (cap. 13).

Le funzionalità dell'area dei dati intermedi riguardano la visualizzazione della forma d'onda e del relativo sonogramma del segnale audio, e la visualizzazione e l'*editing* della segmentazione fonetica (cap. 7) e dell'andamento dei dati di animazione (cap. 8).

Una novità importante, introdotta in questa ultima versione di InterFace, è il meccanismo di sincronizzazione delle rappresentazioni dei parametri con riprese video reali e sintetiche. Con la pressione del *mouse* in un punto delle curve parametriche o del sonogramma, si ottiene il posizionamento sincrono sul *frame* 3D dei *marker* di acquisizione, sul *frame* di faccia sintetizzata, ed anche sul corrispondente fotogramma di un filmato reale (se questo esiste) (Fig. 20, cap. 9).

Per quanto riguarda l'area della sintesi, le Facce Parlanti possono essere attivate sul proprio computer, oppure lanciate in modalità *server*, trasmettendo e ricevendo i dati di animazione in una rete locale oppure sul *Web*, ad uno specifico indirizzo IP (Fig. 2).

La sintesi audio si ottiene ora non solo con il sistema Mbrola (Mbrola Home Page) (Drioli *et alii*, 2003, Drioli *et alii*, 2004), come nelle versioni precedenti del sistema, ma anche con la tecnica SMS (Spectral Modeling Synthesis) (SMS Home Page) (Sommavilla *et alii*, 2005). La rappresentazione dei dati in SMS è intrinsecamente di tipo spettrale, per cui si presta facilmente ad elaborazioni complesse, senza presentare problemi di giunzione fra un difono e il successivo.

## 4. ELABORAZIONE DEI DATI BIMODALI

I dati reali bimodali sono acquisiti da ELITE, un sistema optoelettronico che cattura gli andamenti cinematici di *marker* riflettenti la luce all'infrarosso, e che contemporaneamente campiona l'eventuale segnale audio presente.

Oltre a questi, il sistema accetta nella versione corrente anche i dati acquisiti con EMA (*Electromagnetic Midsagittal Articulometer*) (Zierdt, 1993), che permette di inseguire l'andamento di punti di articolazione vocale non visibili esternamente, per ottenere ad esempio la forma della lingua.

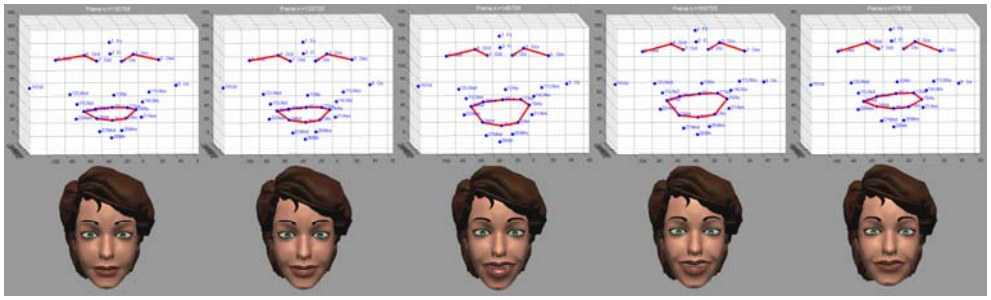

Figura 3a – Fotogrammi 3D di dati reali ricostruiti con Track (in alto), e risintesi (Data-Driven Synthesis) di una faccia esprimente gioia.

I dati audio-visuali possono essere manipolati da tre diverse applicazioni:

- **Track**: permette la ricostruzione 3D delle traiettorie dei *marker* applicati sulla faccia di un soggetto (cap. 5, Fig. 3a, 3b, 7, 10, 11). Questi dati sono poi passati alle altre applicazioni (Optimize, APmanager, visualizzatori vari ed *editor*, ecc.) per le successive fasi di elaborazione. Track consente anche di risintetizzare l'animazione facciale, convertendo queste traiettorie in un flusso di controllo secondo un protocollo voluto (attualmente MPEG-4). Si ottiene in questo modo una tipica ***Data-Driven Synthesis*** (Fig. 3a, 3b) (Damper, 2001; Cosi *et alii*, 2004b).

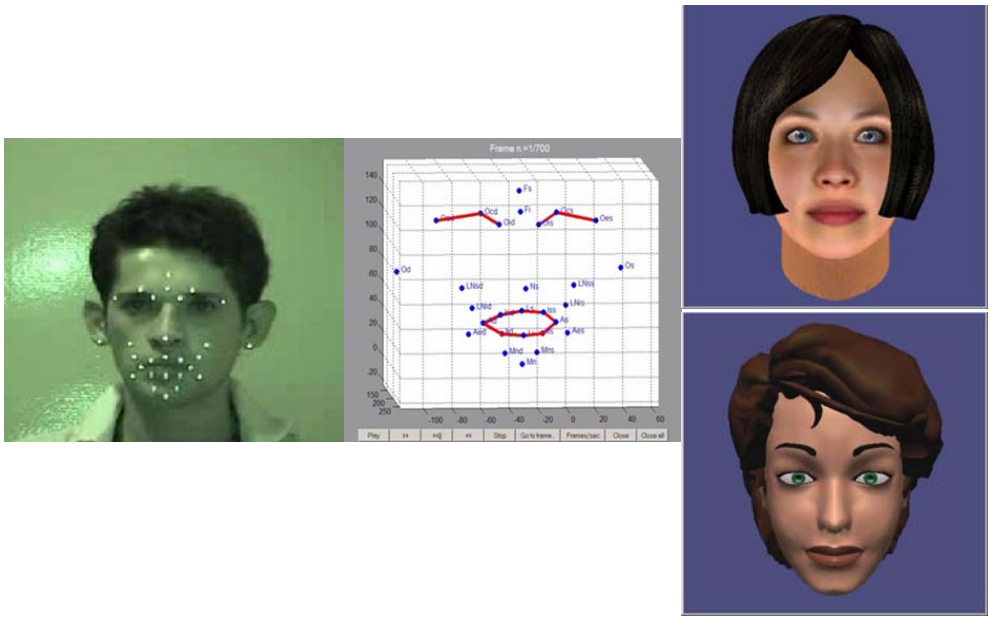

Figura 3b – A sinistra, filmato di un soggetto reale con una configurazione di *marker*. Al centro, sequenza dei punti ricostruiti con Track. A destra, risintesi con Lucia e Greta (*Data-Driven Synthesis*). I filmati sono relativi alla frase: "*…Il fabbro lavora con forza usando il martello e la tenaglia…*"

- **Optimize**: utilizza i dati provenienti da Track per estrarre i coefficienti di articolazione fonetica (cap. 10, Fig. 21-22). I valori di questi coefficienti sono ottimizzati secondo un criterio di minimizzazione dell'errore secondo il modello di Cohen-Massaro (Cohen & Massaro, 1993), che è stato modificato per tener conto della coarticolazione (Perin, 2001; Cosi *et alii*, 2002a, 2002c; 2003a). Questo termine indica il fenomeno di variabilità acustica ed articolatoria delle unità fonetiche, che risultano fortemente dipendenti dal contesto in cui si trovano (Öhman, 1966, 1967; Henke, 1966; Daniloff & Moll, 1973; Bladon & Al-Bamerni, 1976; Bell-Berti & Harris 1981; Al-Bamerni & Blandon, 1982; Keating, 1990; Farnetani & Recasens, 1999). Il modello è stato implementato in **AVengine**, che è il motore per l'animazione da testo scritto (***Text-to-Animation Synthesis***) e da *file* audio (***Wav-to-Animation Synthesis***).
- **APmanager**: consente di stabilire un certo numero di misure fra le traiettorie articolatorie fornite da Track, rispetto a punti, rette o piani di riferimento opportunamente definiti (Fig. 4). APmanager fornisce anche la possibilità di visualizzare e modificare gli andamenti dei parametri voluti, le curve di velocità e accelerazione relative, e di estrarre i punti di massimo e minimo per l'individuazione dei gesti articolatori (cap. 6, Fig. 13-16).

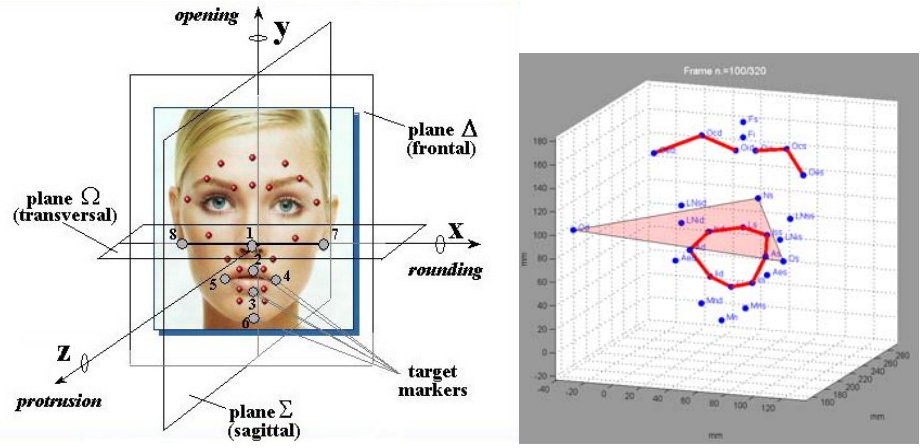


Figura 4 – Piani di riferimento e punti usati in diverse configurazioni, per l'acquisizione dei movimenti articolatori.

## 5. TRACK

Questo programma permette la ricostruzione tridimensionale delle traiettorie di appositi *marker*, e cioè dischetti di plastica applicati sulla faccia di un soggetto e riflettenti la luce infrarossa (Fig. 4-7). In Fig. 4 si vedono i punti di diverse possibili configurazioni: la prima, adottata all'ISTC in passato per misure dei soli movimenti labiali (8 punti grigi), e la seconda, a destra, adottata più recentemente per ottenere un maggior dettaglio articolatorio e per studiare le emozioni (28 punti). Si notano anche i piani di riferimento convenzionali (frontale, trasversale e sagittale) rispetto ai quali si possono definire ed estrarre le misure dei parametri voluti. In Fig. 5 è riportata la posizione assoluta nello spazio dei punti sulla faccia del soggetto che è ripreso, e i punti (speculari) così come risultano proiettati sul piano focale delle due telecamere (vedere anche Fig. 6).

In Fig. 7, si vedono le traiettorie 3D effettivamente ricostruite dei *marker*, nel caso della pronuncia di una frase ("*…Il fabbro lavora con forza usando il martello e la tenaglia…*"), simulando tristezza.

Nella Fig. 3b, si può vedere, a sinistra, un esempio concreto di filmato di un soggetto reale con una configurazione di 28 *marker* applicati alla faccia, mentre pronuncia la stessa frase. Al centro, compare l' animazione relativa dei punti ricostruiti con Track. A destra, infine, si può vedere la risintesi con Lucia e Greta (*Data-Driven Synthesis*).

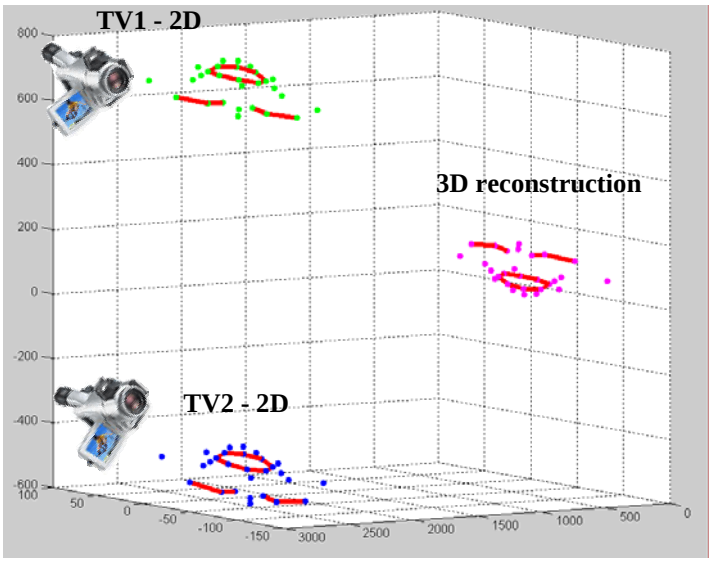


Figura 5 – Track: Ricostruzione 3D (in mm) dei punti reali della faccia (a destra, magenta) dalle immagini 2D catturate dalle telecamere (a sinistra, verde e blu). L'origine delle coordinate è nel parallelepipedo (volume di calibrazione) contenente la faccia originale.

La cattura dei movimenti articolatori avviene, nel nostro caso, con una prima fase di acquisizione dei dati in tempo reale mediante il sistema ELITE. I dati sono acquisiti sotto forma di coordinate *xy* dei *marker*, sul piano focale di un certo numero di telecamere (Fig. 6). L'uso di *marker* passivi ha il grande vantaggio di non ostacolare in nessun modo il movimento degli articolatori della faccia del soggetto, ma presenta anche uno svantaggio non indifferente rispetto all'uso di *marker* attivi (come ad esempio in EMA): il sistema rileva la posizione dei punti, ma non può facilmente risalire alla loro identità, proprio perché questa dovrebbe essere dedotta dalla sola luminosità. Due punti troppo vicini, ad esempio, possono essere interpretati come uno solo, e la posizione rilevata non concorda con nessuno dei due punti originali, poiché il calcolo è basato sul centroide della luminosità dei pixel contigui nell'immagine catturata. I risultati dell'acquisizione possono inoltre essere errati per altri motivi: per la mancanza di punti e/o per la presenza di punti spuri dovuti a riflessi luminosi. È dunque necessaria una seconda fase di post-elaborazione per tentare la corretta identificazione dei punti. Una volta individuati i punti e le loro coordinate 2D, con una tipica procedura di stereofotogrammetria è possibile risalire alla posizione effettiva nello spazio 3D, che non è altro che il punto in cui coincidono le proiezioni passanti attraverso i punti 2D del piano focale e l'obbiettivo della TV relativa (Fig. 5) (Wolf, 1983).

In Fig. 6 si può vedere il fotogramma (il 227° su 741 totali) di due telecamere con la posizione di 28 punti sul piano focale relativi ad una certa configurazione. Le TV devono essere almeno due per ottenere le due rette di proiezione indispensabili per determinare il punto di coincidenza suddetto. Le misure in mm si riferiscono alla distanza dei punti dall'asse focale delle TV.

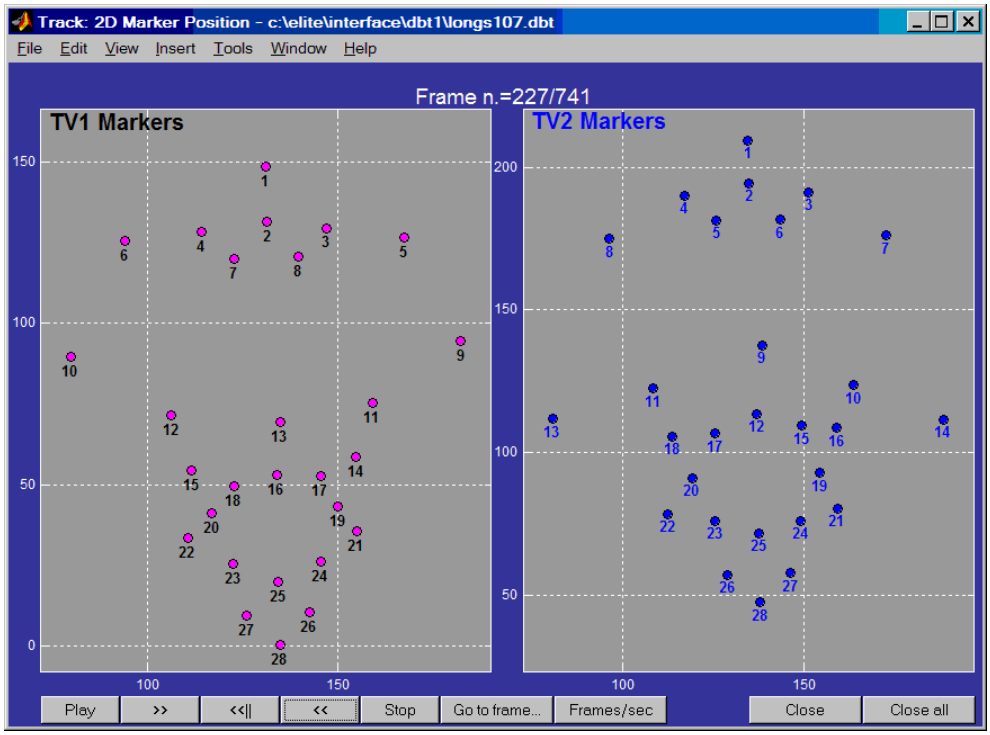


Figura 6 – Track: visualizzazione e animazione dei punti 2D (non ancora identificati) catturati da una faccia esprimente rabbia. In questo fotogramma, sul piano focale delle 2 telecamere, compaiono tutti i 28 *marker* di acquisizione.

Per arrivare a posteriori all'identità dei punti tracciati, nel caso della faccia, come di un oggetto meccanico, è necessario separare il movimento di una delle parti, da quello di roto-traslazione dell'intero sistema. In questo modo, si evita di confondere l'inclinazione in avanti o indietro di tutta la testa, ad esempio, con un moto verso il basso o verso l'alto degli articolatori facciali, che non si sta verificando o che avviene in senso contrario (vedi Fig. 3b). In un corpo rigido, le distanze dei punti da tracciare rimangono costanti. È quindi facile calcolare le componenti di roto-traslazione e il conseguente cambio delle coordinate, che permette di misurare gli spostamenti relativi delle parti volute. In una faccia, ed in particolare nella bocca, queste distanze subiscono delle modifiche notevoli, per cui bisogna disporre almeno di un triangolo di punti indeformabili, su cui si possa basare la trasformazione degli assi cartesiani.

Solo le misure differenziali, calcolate cioè come differenza della posizione reciproca di due punti (ad es. apertura labiale, arrotondamento, ecc.), sono immuni da errori. Le altre risultano affette da un errore tanto più rilevante, quanto maggiore è la deformazione del

triangolo. Anche un triangolo formato dalla punta del naso e dalle due orecchie, che è stato usato in passato, è soggetto a deformazioni. Questo è particolarmente evidente nel caso delle emozioni, che provocano spostamenti sensibili della punta del naso e dei lobi delle orecchie. È necessario, dunque, disporre di un triangolo di punti che non subisca alterazioni per effetti articolatori, come, ad esempio, uno solidale con la calotta cranica.

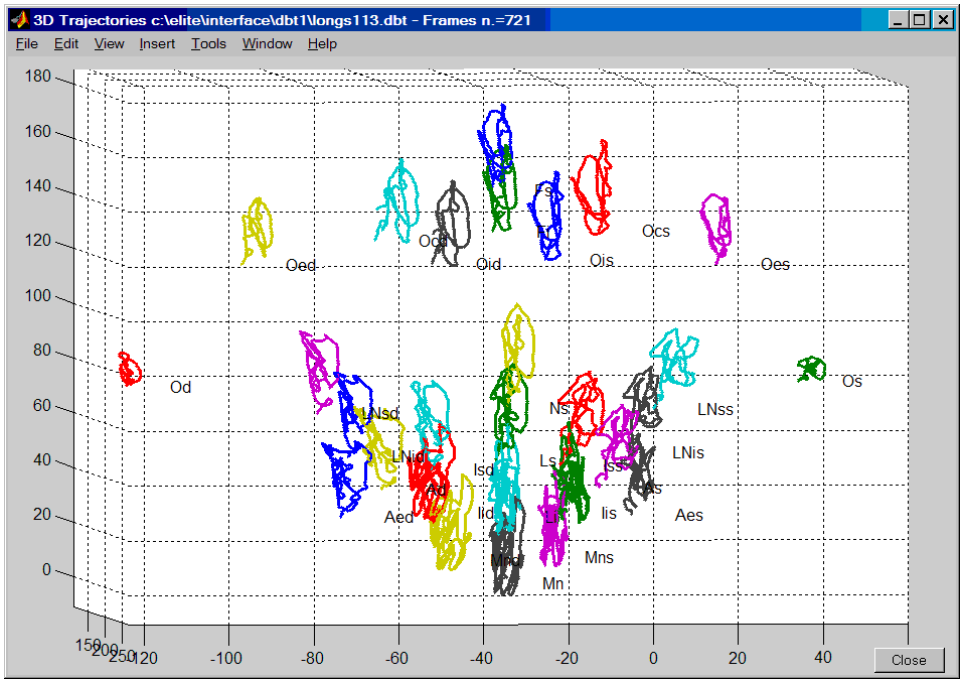


Figura 7 – Track: Traiettorie 3D ricostruite da una faccia esprimente tristezza, mentre pronuncia la frase: "*…Il fabbro lavora con forza usando il martello e la tenaglia…*".

Come era stato anticipato nell'introduzione, il *software* fornito con il sistema ELITE, progettato per il movimento di arti rigidi, si è dimostrato inadeguato nella ricostruzione 3D delle articolazioni e delle espressioni facciali, per la vicinanza e la deformazione della struttura dei *marker* di rilevamento. Un secondo, ma non meno grave, ostacolo è costituito dalla necessità di settare, per ognuno delle migliaia di stimoli acquisiti, una apposita maschera, o modello di riferimento, che stabilisce la corrispondenza fra i punti rilevati e la loro identità, e consente la loro corretta attribuzione alle traiettorie corrispondenti. La eccessiva propensione agli errori del programma fornito con ELITE e la richiesta di un continuo intervento operativo rendevano impossibile affrontare in tempi ragionevoli l'elaborazione di migliaia di *file*.

### *5.1. INNOVAZIONI INTRODOTTE CON TRACK*

Le problematiche viste hanno portato allo sviluppo di un *software* originale detto **Track**.

Le innovazioni introdotte con Track sono le seguenti:

- Il modello di riferimento, a differenza del *software* ELITE, rimane unico per tutta una sessione di lavoro, ovverosia per tutti i *file* di acquisizione, in cui la configurazione dei punti non sia stata modificata (cap. 5.2). Una volta definita la maschera valida per la sessione, l'innesco del processo di riconoscimento delle traiettorie avviene automaticamente per ogni *file*, allineando i punti del primo fotogramma in esame e la maschera con la *Singular Value Decomposition* (SVD) (Soderkvist & Wedin, 1993).
- Anche per risolvere il problema della identificazione dei punti e della deformazione del triangolo di riferimento, è stata usata la SVD. In questo caso la SVD ha il vantaggio di operare intrinsecamente una minimizzazione dell'errore nel calcolo della roto-traslazione.
- La maggior parte del lavoro è stato automatizzato. Track può funzionare sul singolo *file*, oppure su intere *directory*, senza alcuna necessità di interventi manuali. Si può ovviamente in una fase successiva di controllo correggere eventuale errori.
- Nella generazione del *FAP-stream* necessario per l'animazione, la corrispondenza fra i punti di acquisizione e i punti dello standard MPEG-4 è completamente riconfigurabile (cap. 5.4). Questo implica la possibilità di adottare un qualsiasi altro protocollo si voglia utilizzare per l'animazione.
- Il *FAP-stream* prodotto tiene conto della roto-traslazione della testa e dei fattori di scala della testa che si vuole animare, il che permette una corretta ***Data-Driven Synthesis*** di un qualsiasi agente MPEG-compatibile.

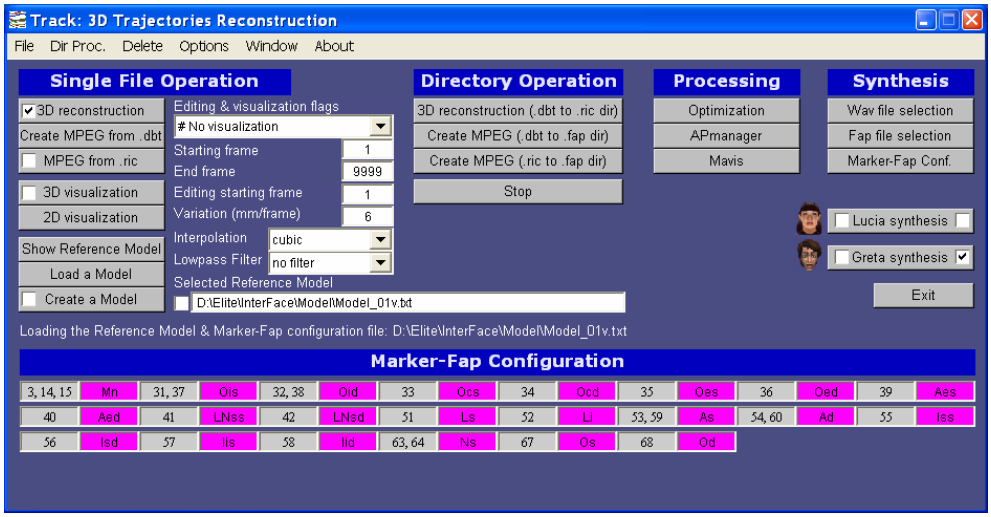


Figura 8 – Track: Schermata principale del programma.

In Fig. 8 si può vedere la pagina principale del programma Track con le principali aree funzionali: ***Operazioni su file singoli***, ***Operazioni su directory***, ***Elaborazione***, ***Sintesi***.

L'area sulla sinistra riguarda le operazioni compiute su un singolo *file* di dati, e comprende la ricostruzione 3D di cui abbiamo parlato in precedenza, la conversione a dati MPEG-compatibili, la visualizzazione e l'*editing* delle traiettorie bi- e tri-dimensionali dei

*marker*, e la costruzione del modello di riferimento, che, come detto, è unico per tutta la sessione di lavoro (cap. 5.2).

L'area in basso riporta la corrispondenza fra i parametri di animazione FAP e le traiettorie dei *marker* che si vogliono controllare, in vigore in quel momento. La presenza di più di un numero identificativo per i FAP significa che il controllo avviene lungo diversi assi cartesiani. Ad esempio, la prima casella a sinistra *Mn*, relativa alla mandibola contiene il movimento su tutti e tre gli assi di riferimento. Un clic sui pulsanti relativi permette di riconfigurare a piacere la corrispondenza *marker*-FAP (cap. 5.4).

### *5.2. TRACK: CONFIGURAZIONE DEL MODELLO DI RIFERIMENTO*

Il modello di riferimento permette di identificare i punti rilevati dal sistema ELITE. È necessario impostarlo solo *una tantum* e, come detto, rimane unico per tutta una sessione di lavoro, ovverosia per tutti i *file* di acquisizione, in cui la collocazione dei *maker* sulla faccia del soggetto non sia stata alterata.

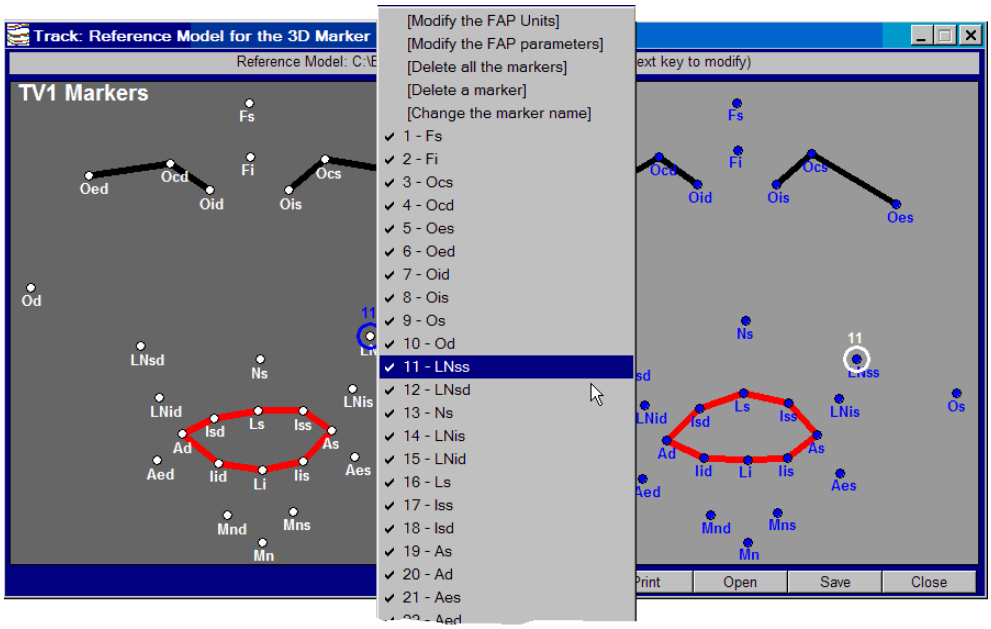


Figura 9 – Track: Modello di riferimento per la ricostruzione delle traiettorie articolatorie.

La configurazione è fatta attribuendo una etichetta ad ognuno dei punti, ed eventualmente collegandoli con linee colorate in modo da rendere l'animazione più fruibile (Fig. 9), ed è poi salvata in un *file* opportuno. La posizione dei punti della maschera può essere letta direttamente da uno qualsiasi dei *file* di acquisizione corrispondenti a quella sessione, o anche, volendo, impostata manualmente nel *file* di configurazione.

L'identificazione avviene confrontando i punti del modello di riferimento con quelli del fotogramma corrente nel *file* in analisi, dopo che la SVD ha permesso la roto-traslazione dei dati in modo da farli combaciare con il riferimento. A questo punto l'identità di un punto incognito deriva dalla minima distanza con uno dei punti noti (*nearest neighbour*

*association rule*). Una volta soddisfatti certi criteri (che non ci siano errori di doppie assegnazione e che compaiono tutti i *marker*), il fotogramma dei dati correnti diventa la nuova maschera, o nuovo modello di riferimento, adattandosi progressivamente anche a grandi deformazioni temporali della struttura dei punti.

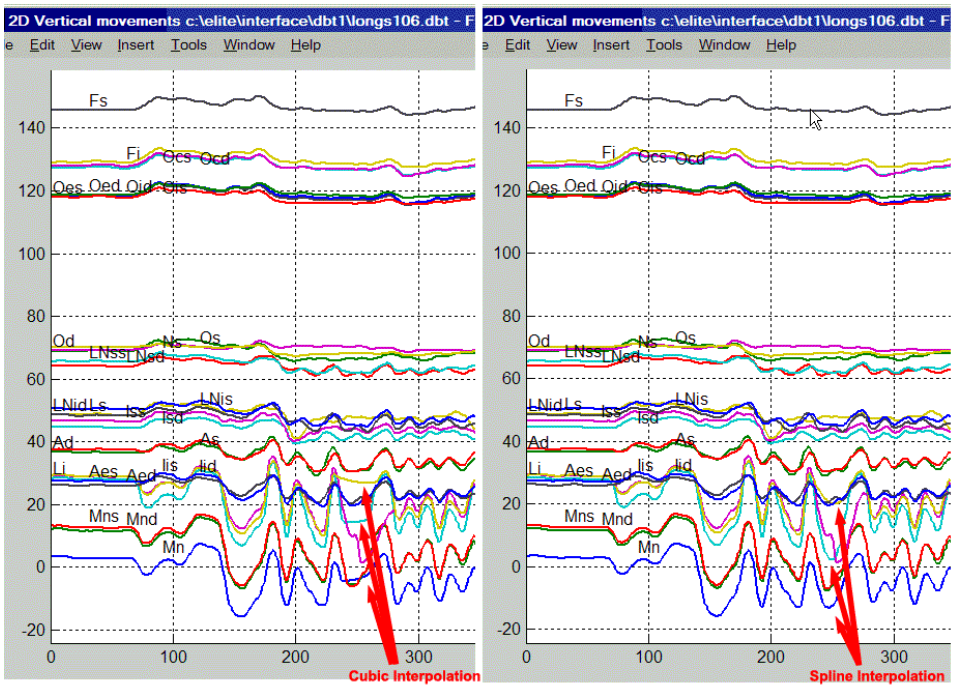


Figura 10 – Track: Tipi di interpolazione nella ricostruzione di traiettorie articolatorie.

### *5.3. TRACK: RICOSTRUZIONE DELLE TRAIETTORIE DEI MARKER*

Nella finestra di Fig. 8 si possono vedere a sinistra i pulsanti che permettono la ricostruzione 3D delle traiettorie, l'animazione 2D e 3D fotogramma per fotogramma, e la preparazione della maschera di controllo.

Nella colonna a fianco, si impostano i parametri di controllo della ricostruzione delle traiettorie articolatorie (tipi di visualizzazione, fotogramma iniziale e finale di analisi, eventuale *frame* da cui iniziare l'*editing* manuale delle traiettorie, variazione massima per l'identificazione dei punti, tipo di interpolazione, tipo di filtraggio, ecc.).

Uno di questi parametri è la massima variazione tollerabile (in mm) dello spostamento di un punto da un fotogramma all'altro. Questo permette di decidere se la nuova posizione può essere considerata accettabile, o fuori scala. Il parametro va aggiustato a seconda degli stimoli registrati, aumentando l'escursione possibile, se si tratta di emozioni, ad esempio, in cui i movimenti risultano molto ampi.

Altri parametri permettono di scegliere se interpolare o no i dati, e quale tipo di interpolazione (cubica, *spline*, lineare) si voglia applicare. In caso di una traiettoria con pochi punti mancanti il risultato migliore è dato dalla *spline* (vedi Fig. 10).

Si può anche decidere se applicare un filtraggio passabasso o no, tenuto conto che nella fase di estrazione dei veri e propri parametri articolatori il filtraggio sarà comunque forzatamente imposto (cap. 6). In Fig. 11 si vede il risultato del filtraggio con un filtro Lambda (*Linear-Phase Autoregressive Model-Based Derivative Assessment Algorithm*) appositamente studiato per questo tipo di applicazioni (D'Amico & Ferrigno, 1990, 1992).

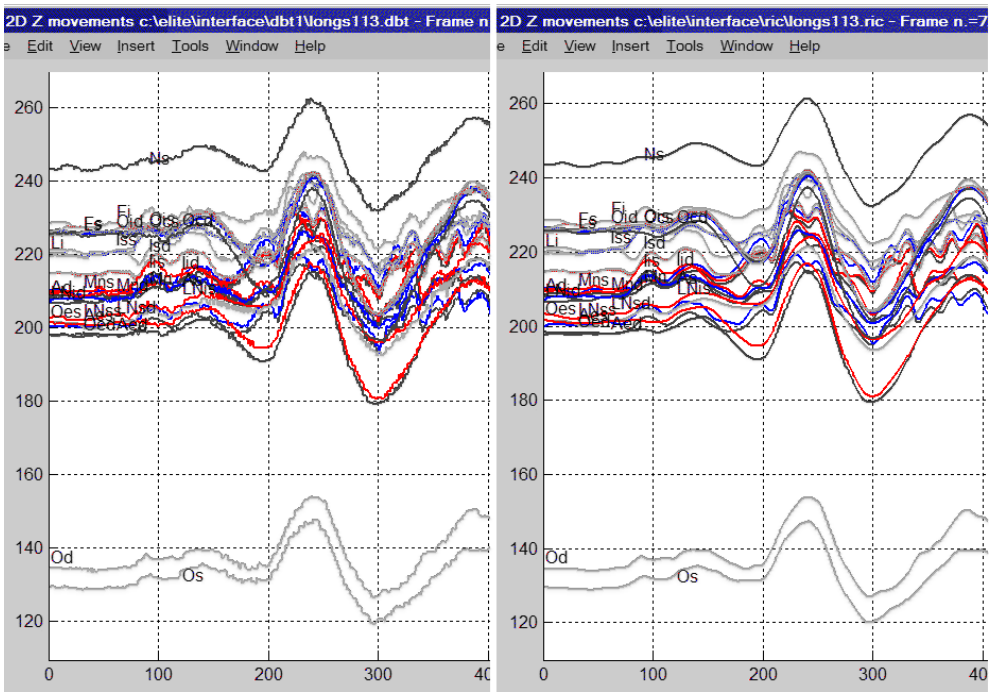


Figura 11 – Track: a sinistra traiettorie di 28 marker della faccia rispetto all'asse Z (antero-posteriore) caratterizzate da una forte componente aleatoria. A destra, le stesse traiettorie dopo il filtraggio *Lambda* (*Linear-Phase Autoregressive Model-Based Derivative Assessment Algorithm*).

### *5.4. TRACK: CONFIGURAZIONE DELLA CORRISPONDENZA MARKER-FAP*

Per l'animazione si è adottato il protocollo MPEG-4, che offre vari vantaggi fra cui il fatto che è uno standard di accesso in rete alle *Talking Heads* (Fig. 2). MPEG-4 ha la possibilità di:

- Animazione a distanza di una faccia 3D costruita su un reticolo (*mesh*) poligonale di punti con un flusso di dati detti FAPs (*Facial Animation Parameters*), generati secondo lo standard MPEG-4 (mpeg.telecomitalialab.com/standards/MPEG-4), (Doenges, 1997; Lavagetto & Pockaj, 1999).
- Le Facce Parlanti dell'ISTC, Lucia e Greta, utilizzano per l'animazione un approccio pseudo-muscolare, nel quale la contrazione del muscolo è ottenuta attraverso la deformazione della *mesh* attorno a punti particolari (*feature points*), che corrispondono all'attaccatura dei muscoli della faccia. Ai FAPs, prima definiti, corrispondono minime azioni facciali.

- Adattamento dei FAPs alla conformazione di una particolare faccia mediante i *Facial Animation Parameter Units* (FAPU).
- Definizione eventuale dell'intera struttura della faccia con i *Facial Definition Parameters* (FDPs), con cui imporre fra l'altro la tipologia maschio o femmina, vecchio o giovane, e altri particolari come occhiali, cappello, ecc.

Per gli scopi di ricerca fonetica e nel campo delle emozioni, è molto importante disporre di mezzi che permettano la rapida configurazione e manipolazione di questi parametri. È per questo motivo che si è creato un apposito *tool* per stabilire la corrispondenza fra i punti di acquisizione con ELITE e quelli dello standard MPEG-4 (Fig. 12).

In Fig. 12 si può vedere sulla sinistra i punti dello standard MPEG-4, mentre sulla destra compaiono i punti della configurazione usata per l'acquisizione. L'associazione fra gli uni e gli altri è completamente riconfigurabile con un menu contestuale. Questo implica anche la possibilità di modificare lo stesso standard MPEG-4 e/o di adottare un qualsiasi altro protocollo di animazione.

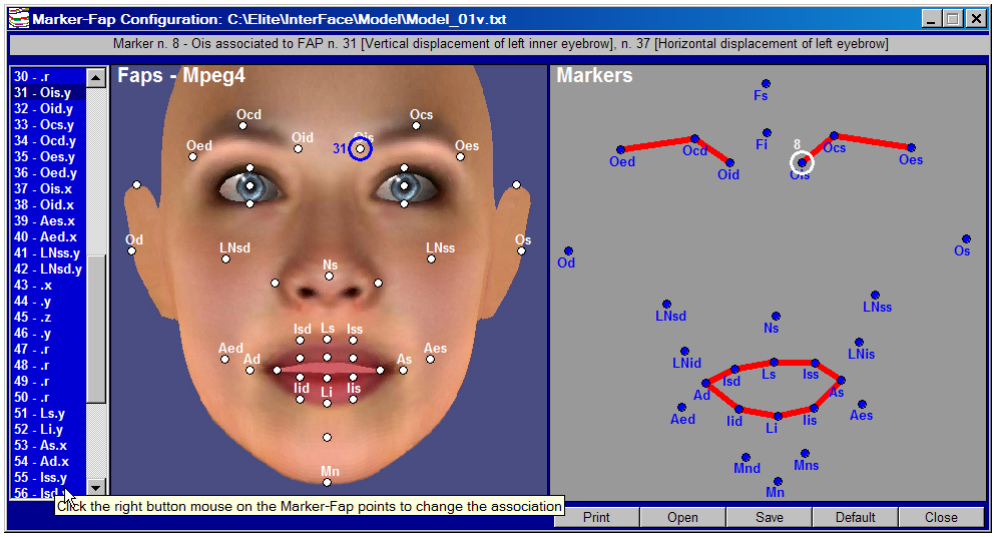


Figura 12 – Track: Configurazione della corrispondenza fra *marker* e punti MPEG-4.

## 6. ESTRAZIONE ED *EDITING* DEI PARAMETRI ARTICOLATORI

Le grandezze articolatorie significative dal punto di vista fonetico e delle emozioni sono ricavate dai dati fisici delle traiettorie nello spazio 3D, depurati dalle componenti di roto-traslazione della testa (vedi inizio del cap. 4).

Per far questo, si è creato un programma **APmanager** (***Articulatory Parameter Manager***) che permette di determinare una qualsiasi misura voluta del tipo: distanza da *marker* a *marker*, distanza da *marker* ad una retta prestabilita, distanza da *marker* da un piano di riferimento (Fig. 4), angolo fra rette e/o piani voluti.

In Fig. 13 si può vedere la finestra principale del programma con un *file* di dati precedentemente elaborato da Track, il *database* contenente l'insieme dei parametri definiti fino a quel momento, e nel riquadro arancio il *set* dei parametri espressamente associati a questo *file*.

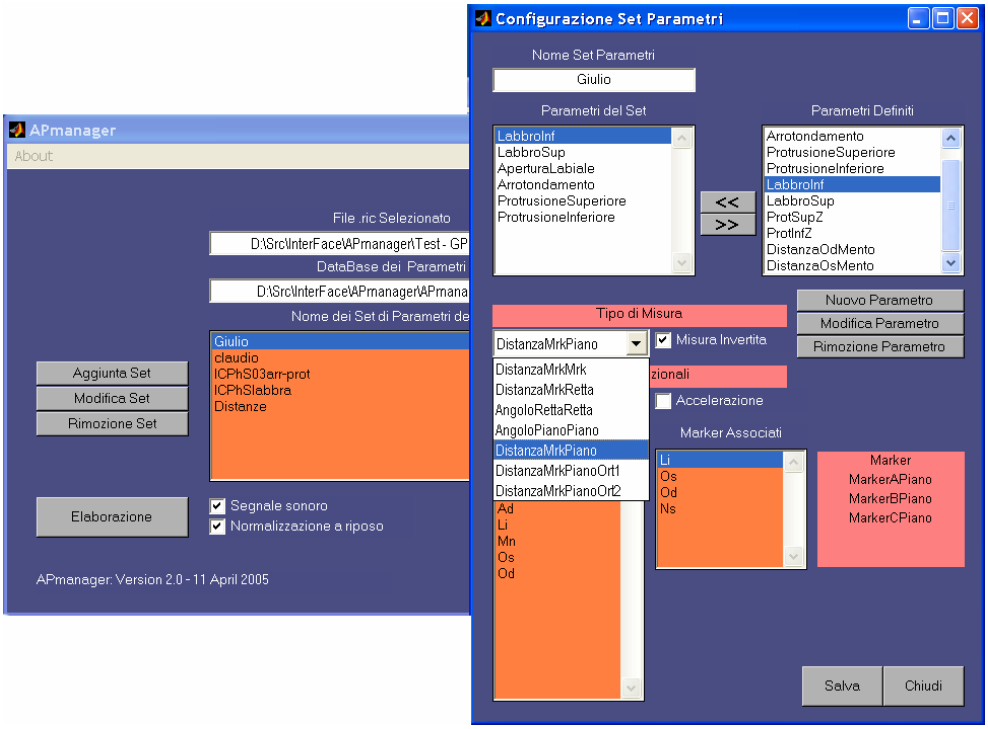


Figura 13 – APmanager: finestra principale e, sovrapposta, pannello di configurazione di un *set* di parametri con le misure ricavate da 8 *marker*.

Sovrapposto a questa, compare la finestra di configurazione di uno dei *set* di misure (in questo caso il *set* di nome "Giulio"), con i parametri definiti in questo particolare caso. A destra, compaiono tutte le misure del *database* che già sono state definite (*Arrotondamento*, *ProtrusioneSuperiore*, *ProtrusioneInferiore*, *LabbroInf*, *LabbroSup*, *ProtSupZ*, *ProtInfZ*,

*DistanzaOdMento*, *DistanzaOsMento*, ecc.). A sinistra, sotto il titolo "*Parametri del Set*", si possono vedere, le misure effettivamente previste per il *file* in questione: *LabbroInf*, *LabbroSup*, *AperturaLabiale*, *Arrotondamento*, *ProtrusioneSuperiore*, *ProtrusioneInferiore*, che corrispondono al movimento del labbro inferiore e superiore, all'apertura labiale, all'arrotondamento, alla protrusione inferiore e superiore delle labbra.

Sotto il titolo "*Tipo di Misura*", si può notare la definizione di una delle misure, *LabbroInf* (spostamento verticale del labbro inferiore), come la distanza *DistanzaMrkPiano* del *marker* Li (labbro inferiore) dal piano contenente i marker Os, Od, Ns, corrispondenti al naso e ai lobi delle orecchie (Fig. 14).

Nella finestra di configurazione delle misure, può essere selezionata la misura con segno invertito, e le curve di velocità ed accelerazione associate alle varie misure.

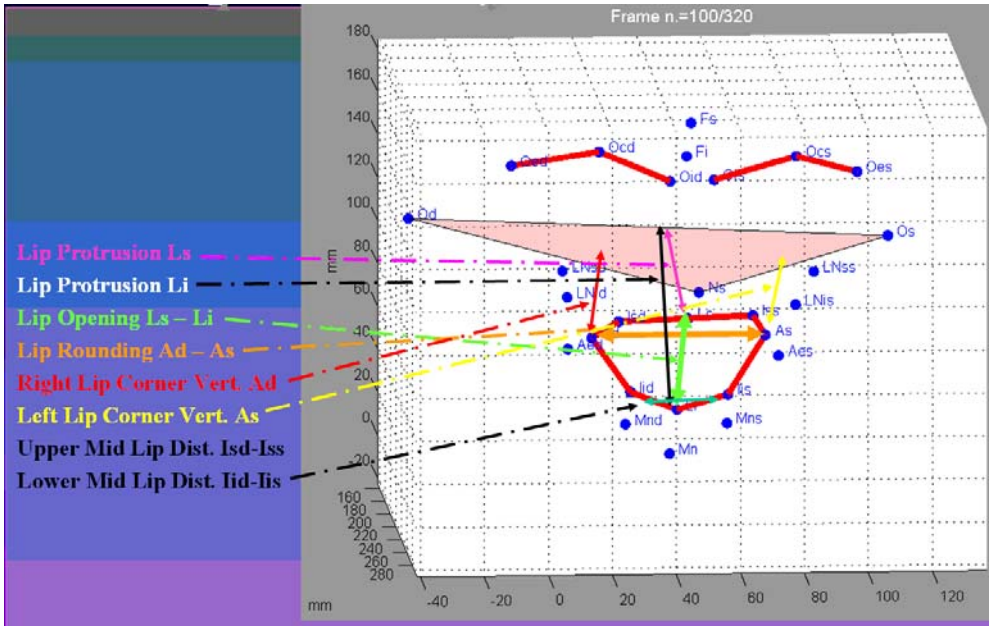


Figura 14 – APmanager: misure di interesse fonetico-fonologico.

I parametri tradizionalmente usati all'ISTC per una descrizione fonetico-fonologica sono i seguenti (Fig. 14):

1. Apertura Labiale (LO)
2. Spostamento Verticale del Labbro Superiore (UL)
3. Spostamento Verticale del Labbro Inferiore (LL)
4. Arrotondamento Labiale (LR)
5. Protrusione del Labbro Superiore (ULP)
6. Protrusione del Labbro Inferiore (LLP)

All'ISTC, ai parametri suddetti sono stati aggiunti i seguenti per lo studio delle emozioni (Fig. 15):

7. Spostamento Orizzontale dell'Angolo Sinistro delle Labbra (LCX)
8. Spostamento Verticale dell'Angolo Sinistro delle Labbra (LCY)

9. Spostamento Orizzontale dell'Angolo Destro delle Labbra (RCX)
10. Spostamento Verticale dell'Angolo Destro delle Labbra (RCY)
11. Asimmetria Orizzontale delle Labbra (ASYMX)
12. Asimmetria Verticale delle Labbra (ASYMY)
13. Aggrottamento della Fronte 1 (distanza fra punti interni sopracciglia)
14. Aggrottamento della Fronte 2 (distanza fra punti centrali sopracciglia)
15. Corrugamento della Fronte
16. Spostamento Verticale dell'Angolo Interno Destro delle Sopracciglia
17. Spostamento Verticale dell'Angolo Interno Sinistro delle Sopracciglia
18. Spostamento Verticale del Punto Centrale Destro delle Sopracciglia
19. Spostamento Verticale del Punto Centrale Sinistro delle Sopracciglia

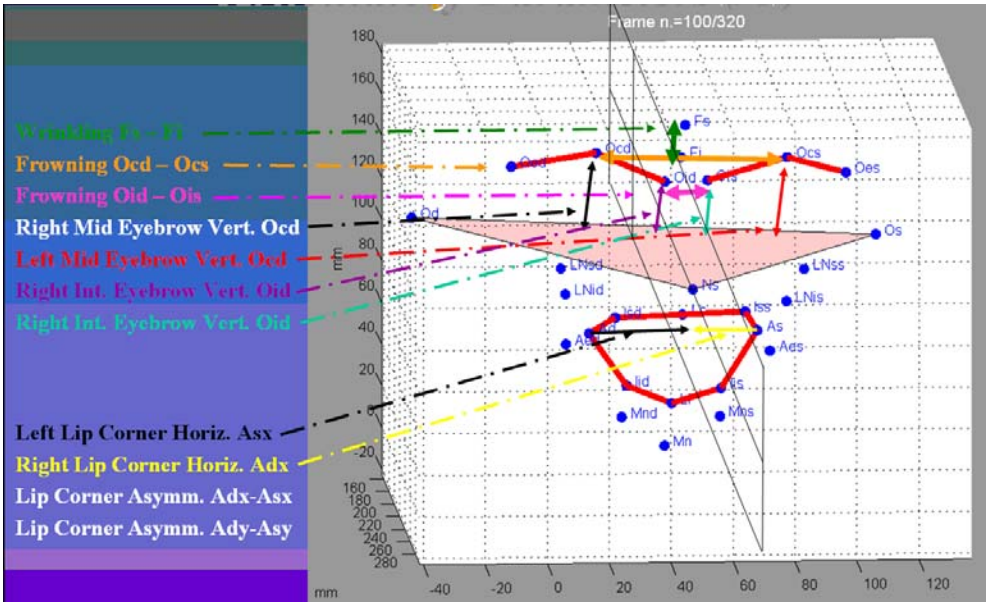


Figura 15 – APmanager: misure definite all'ISTC per lo studio delle emozioni.

Nella corrente versione di InterFace, è stata sostituita la visualizzazione dei parametri articolatori fatta con una applicazione esterna (*Mavis - Multiple Articulator VISualizer,* di Mark Tiede), poiché non prevedeva il sonogramma del segnale audio.

È stata realizzata in questa versione una apposita finestra di visualizzazione, in cui compaiono i parametri in sincronia con la forma d'onda e il sonogramma (quando ovviamente esiste il relativo audio) (Fig. 16).

Con il tasto destro contestuale si può ottenere le curve di velocità e accelerazione relative (che come si è visto possono essere ricavate anche da APmanager). È stata fornita anche la possibilità di *editing* delle curve.

Si è cercato di metter in evidenza i gesti articolatori, contrassegnando i punti di massimo e minimo degli andamenti temporali e i punti dove la velocità decresce o supera una certa soglia percentuale significativa fissata dall'utente. Nel grafico temporale dell'apertura della labbra (*Apertura*Lab) di Fig. 16, la curva di velocità è individuata dal

colore magenta. Si può vedere come in corrispondenza ai punti di massimo (triangolo rosso) o minimo (triangolo verde) dell'andamento di ampiezza della curva (colore blu), la velocità si annulla.

Supponendo di aver fissato, ad esempio, al 15% del massimo locale di velocità la soglia significativa, si può vedere in Fig. 16 il punto dove comincia effettivamente il gesto di apertura delle labbra nell'articolare una /a/ di /aba/ (prima freccia rossa a sinistra successiva al minimo di apertura, a circa 1.62 s). Il gesto articolatorio si conclude, secondo questa convenzione, nel punto contrassegnato dalla freccia verde successiva a destra, precedente il massimo dell'apertura, a circa 1.77 s.

I parametri possono essere modificati semplicemente cliccando sulla curva desiderata e trascinando il *mouse* in modo da disegnare l'andamento temporale voluto.

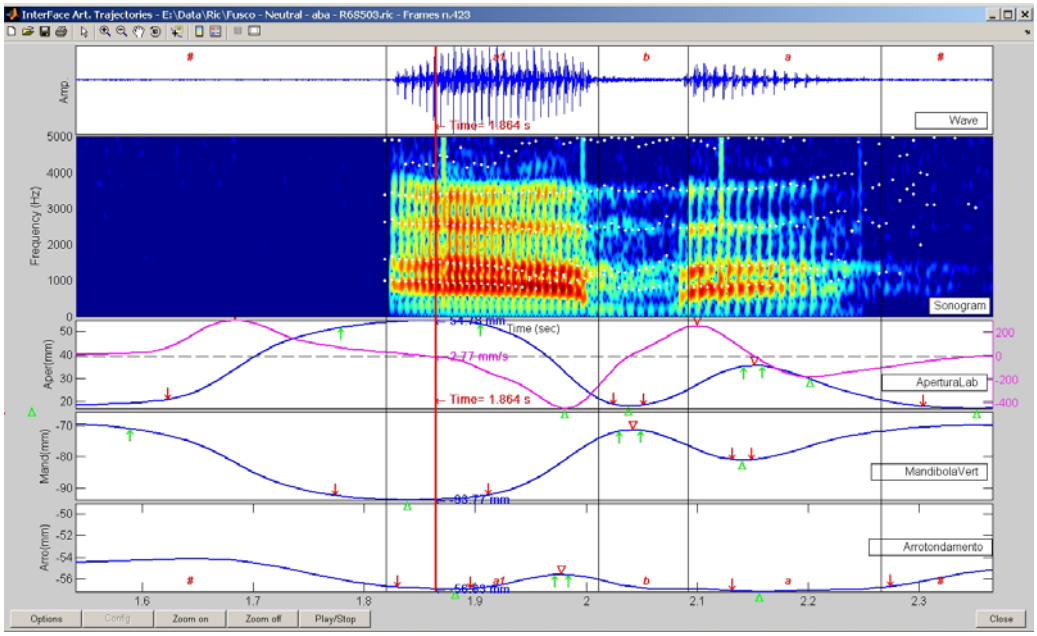


Figura 16 – APmanager: *editing* dei parametri articolatori. Si possono vedere i punti di massimo (triangolo rosso) e minimo (triangolo verde) degli andamenti temporali e della velocità (in magenta). Le frecce rosse e verdi individuano i punti di inizio e fine dei gesti articolatori (compresi entro una certa soglia). Notare all'inizio l'anticipo dei movimenti articolatori rispetto all'emissione vocale.

## 7. SEGMENTAZIONE FONETICA

La segmentazione e il *labelling* dei parametri articolatori ed acustici in unità significative rivestono un'importanza particolare in campo linguistico, specialmente per la formulazione di modelli di co-produzione del parlato (Magno *et alii*, 2001, Magno *et alii*, 2003, Magno *et alii*, 2004).

In InterFace, la segmentazione può essere ottenuta in vari modi:

Nella sintesi da puro testo o da testo XML, la trascrizione fonetica è contestualmente creata dal *parser* di Festival per la stringa da sintetizzare, assieme alle relative durate, essendo, in effetti, proprio questi i dati che servono a Mbrola o SMS per la sintesi voluta.

La Fig. 17 mostra un esempio di sonogramma della frase "(belli)ssima giornata" sintetizzata con Festival-Mbrola, dove sono evidenti i confini delle varie unità fonetiche. Oltre al sonogramma (in basso), compaiono la forma d'onda (in alto), l'andamento del *pitch* in Hz. (sopra al sonogramma), una sezione spettrale con posizione delle formanti a 3.08 s (a sinistra), ed infine una rappresentazione spettrale 3D, detta *Waterfall* (sulla destra).

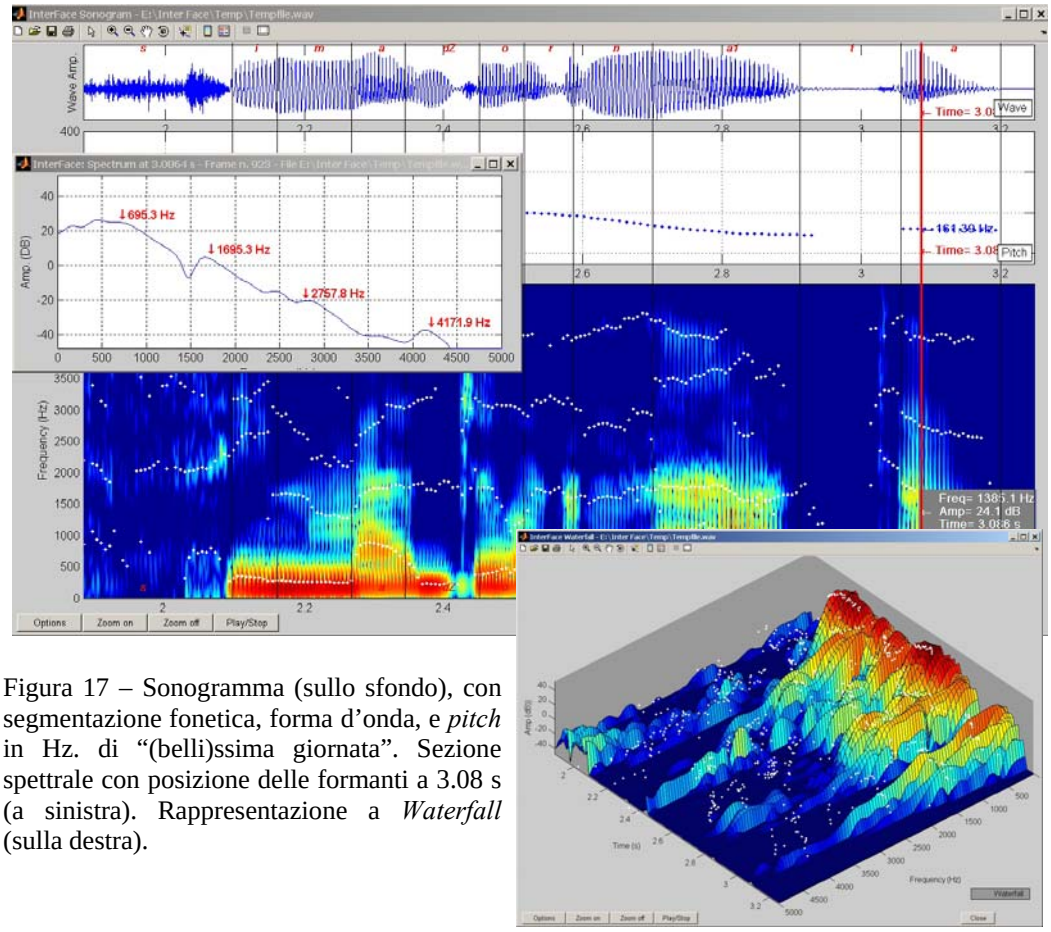


Figura 17 – Sonogramma (sullo sfondo), con segmentazione fonetica, forma d'onda, e *pitch* in Hz. di "(belli)ssima giornata". Sezione spettrale con posizione delle formanti a 3.08 s (a sinistra). Rappresentazione a *Waterfall* (sulla destra).

Un'altra strada per ottenere la segmentazione è quella di partire da un *file* di parlato preesistente. Si fornisce il *file* ad un riconoscitore integrato in InterFace nella funzione *Wav-to-Animation*, per ottenere la sequenza testuale e fonetica relativa. Il *file* prodotto può essere editato nella finestra principale di InterFace (*Phon Edit*). La segmentazione relativa compare nelle finestre del sonogramma come si è visto in Fig. 16 e 17, dove può essere modificata dall'utente. Questa funzionalità richiede l'installazione di uno dei seguenti *software*:

- ***OGI Toolkit***: OGI - CSLU toolkit - OGI School of Science & Engineering - Center for S poken Languages Understanding.
- ***Sonic***: http://cslr.colorado.edu/beginweb/speech_recognition/sonic.html

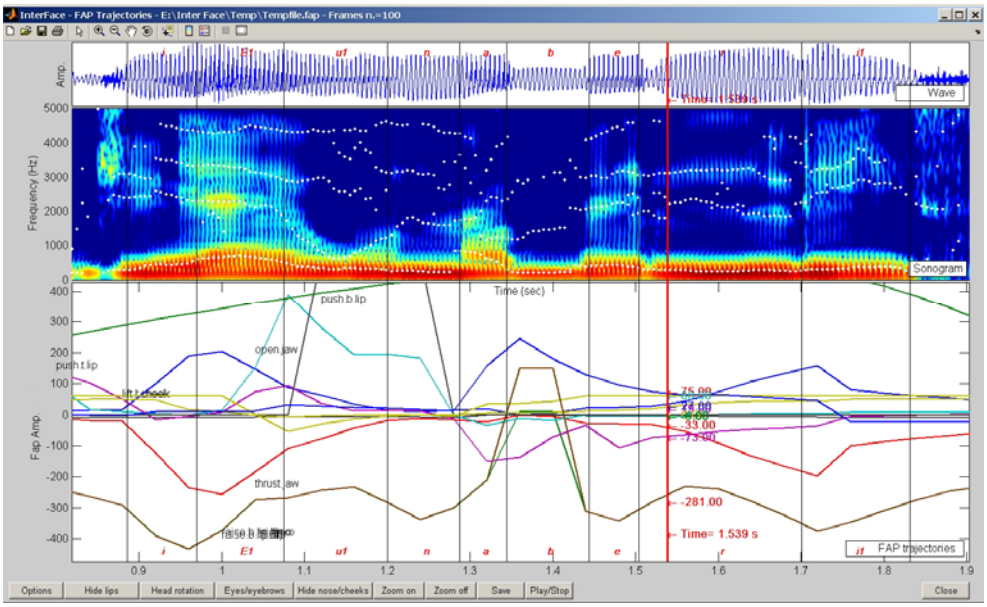


Figura 18 – Display e *editing* FAPs.

## 8. *EDITING* DEI FAPS

Come si è visto nel cap. 5.4, la sintesi delle Facce Parlanti in InterFace è ottenuta con un flusso tempo-variante di dati che producono le deformazioni volute della struttura poligonale dell'agente animato. Si è anche accennato ai *Facial Definition Parameters* (FDPs) che servono alla configurazione della struttura stessa, e ai *Facial Animation Parameter Units* (FAPU) che consentono l'adattamento di un generico FAP-stream alla conformazione di una particolare faccia.

L'andamento dei FAPs può essere visualizzato nella finestra principale di InterFace (pulsante ***Fap Edit***) in sincrono con l'eventuale forma d'onda e con il sonogramma. Alcuni tasti permettono di nascondere e far comparire i gruppi di parametri associati alle labbra, agli occhi e alle sopracciglia, al naso e alle guance, e alla rotazione della testa (Fig. 18).

I FAPs possono essere modificati semplicemente cliccando sulla curva voluta e disegnando con il *mouse* l'andamento desiderato.

Per facilitare il test di un singolo parametro o un gruppo di parametri, è stato inserito nella schermata principale di InterFace un tasto ***Fap on/off***, per accedere ad una finestra, dove si può attivare o disattivare le componenti del FAP-stream desiderate (Fig. 19).

Sulla faccia di Fig. 19, le frecce indicano quale siano i possibili movimenti del punto relativo (orizzontale, verticale, antero-posteriore). I colore indica se quel parametro è correntemente attivo (color verde), inibito (color rosso), oppure non sia stato definito (color grigio). L'attivazione/disattivazione può avvenire cliccando su uno dei punti, oppure da un apposito menu contestuale, oppure anche con appositi pulsanti che agiscono su un gruppo di parametri (labbra, fronte, mandibola, ecc.)

Nelle due colonne a sinistra compaiono i 68 FAPs dello standard MPEG-4, con il loro nome e la direzione secondo la quale agiscono (colonna a sinistra). Ad esempio, il 14° e 15° FAP in questo caso sono effettivamente definiti ed attivi (color verde sulle frecce della faccia), e controllano il movimento lungo l'asse Z (antero-posteriore) e l'asse X (orizzontale). La terza componente attiva corrisponde al FAP 3 (non visibile nella figura).

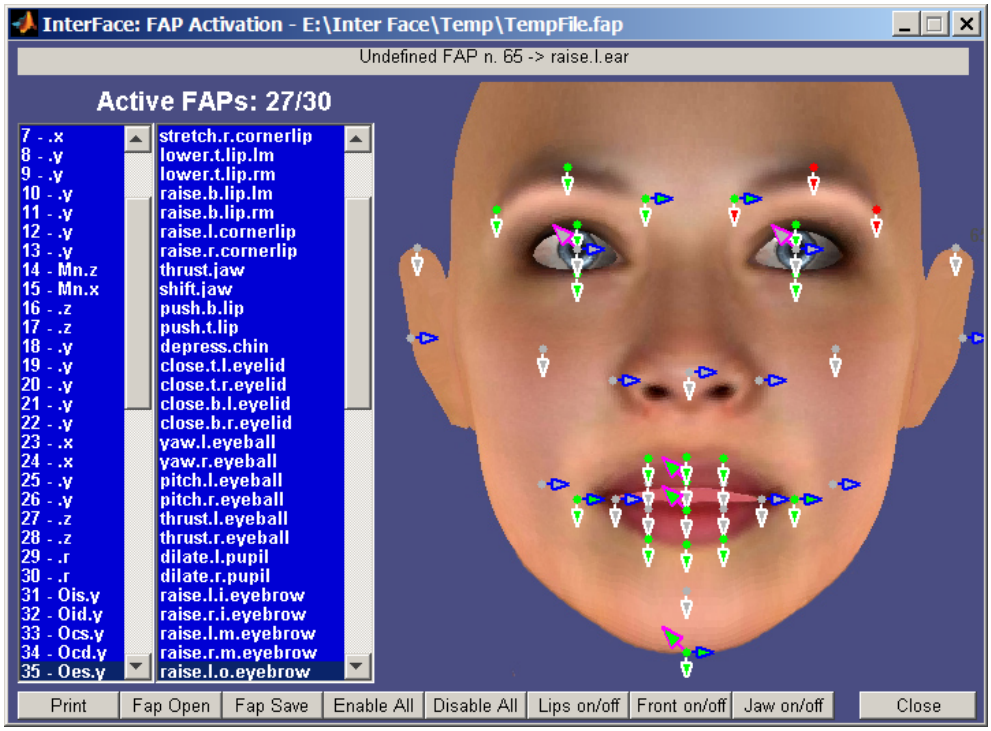


Figura 19 – Finestra per attivare/disattivare le componenti del FAP-stream che animano la Faccia Parlante. In verde sono i FAP attualmente attivi, in rosso quelli inibiti, in grigio quelli non controllati in questa sequenza animata.

## 9. SINCRONIZZAZIONE AUDIO-VISUALE

Come si è detto nell'introduzione, il parlato è inerentemente una comunicazione multimodale, in cui la decodifica del messaggio passa tanto attraverso il canale acustico che quello visivo. La comunicazione ha una natura multimodale anche perché veicola non solo le informazioni linguistiche necessarie per la comprensione reciproca, ma anche informazioni extralinguistiche, che hanno una importanza notevole dal punto di vista relazionale e sociale. Questo flusso di informazioni è strettamente connesso alla nostra costituzione fisica, agli atteggiamenti, alle idiosincrasie, all'umore, allo stato d'animo e alle emozioni, e interessa sia le caratteristiche acustiche (prosodia, intonazione, timbro, ecc.) della voce, che gli aspetti articolatori e visivi della faccia (e del corpo in generale).

Gli sviluppi recenti nel campo della multimodalità hanno coinvolto diversi settori disciplinari dalla linguistica, all'acustica, alle telecomunicazioni, al riconoscimento automatico del parlato, alla psicologia e alla neurofisiologia.

L'esigenza principale, che hanno portato queste ricerche, è quella di registrare e elaborare segnali articolatori e acustici in sincronia, come si è visto nei capitoli precedenti, ma anche di segmentare, indicizzare ed etichettare filmati reali (Kipp, 2004)).

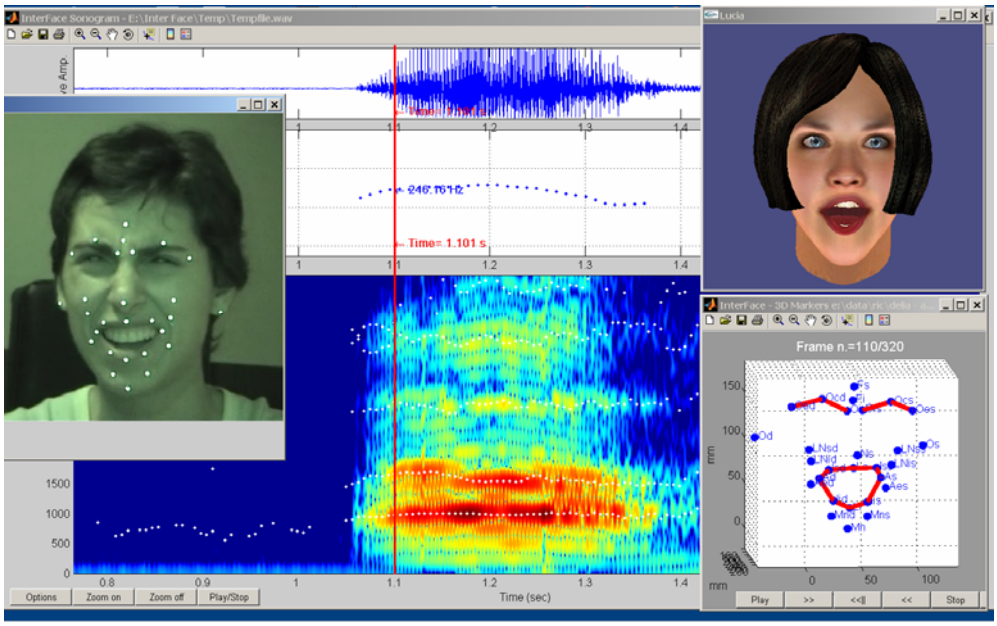


Figura 20 – Sincronizzazione dal sonogramma di un vocalizzo /a/ esprimente rabbia, sul *frame* 3D n. 110 dei *marker* di acquisizione e della faccia sintetizzata, e sul corrispondente fotogramma del filmato reale.

Nella corrente versione di InterFace si è introdotto un meccanismo di sincronizzazione non solo delle rappresentazioni dei parametri audio-visuali, ma anche dei video relativi provenienti tanto da una ripresa reale quanto dalla simulazione con un agente virtuale.

Un clic con il *mouse* sul punto desiderato delle curve articolatorie o sul sonogramma (ad esempio a 1.1 s di Fig. 20) permette la sincronia con:

- con il *frame* 3D della ricostruzione dei *marker* (Fig. 20, in basso a destra)
- del *frame* della faccia sintetizzata (Fig. 20, in alto a destra)
- anche sul corrispondente fotogramma di un filmato reale (Fig. 20, a sinistra).

## 10. OTTIMIZZAZIONE DEI PARAMETRI ARTICOLATORI

È stato sviluppato un apposito modello di coarticolazione fonetica e un *software*, **Optimize**, che estrae i coefficienti relativi dai dati reali provenienti da Track.

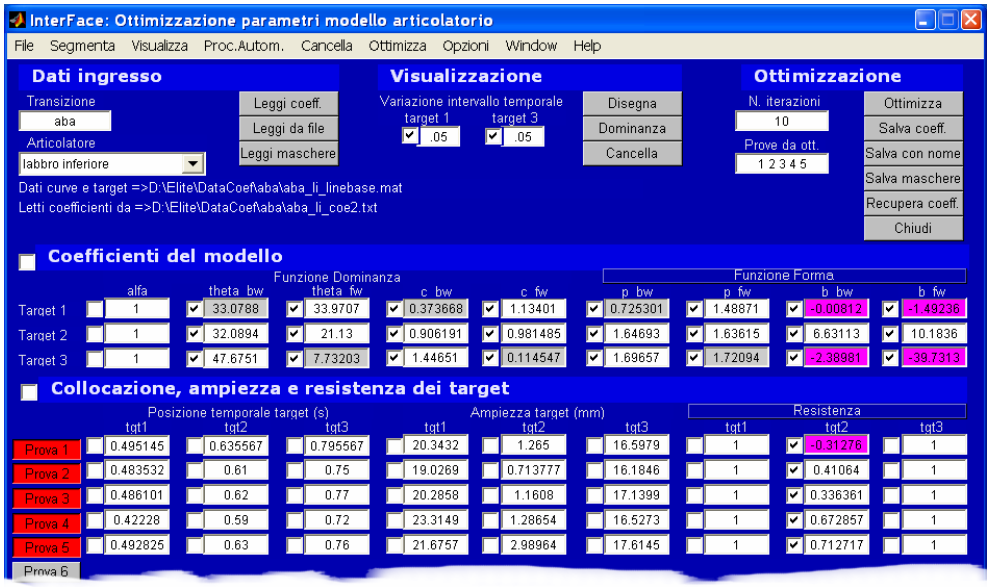


Figura 21 – Optimize: schermata principale con i parametri di ottimizzazione dell'articolazione del labbro inferiore per 5 prove di una transizione /aba/.

Come è stato accennato in precedenza, la coarticolazione è dovuta alla reciproca influenza dei movimenti articolatori durante la produzione del parlato, ed è responsabile della grande variabilità fonetica, che si verifica nelle lingue e dialetti. Questo fenomeno, di per sé stesso complesso e difficile da studiare, è complicato dall'influenza esercitata delle emozioni sull'articolazione fonetica.

Fra i molti modelli proposti in letteratura (Kozhevnikov & Chistovich, 1965; Öhman, 1966, 1967; Chomsky & Halle, 1968; Henke, 1966; Daniloff & Moll, 1973; Bladon & Al-Bamerni, 1976; Bell-Berti & Harris, 1981; Al-Bamerni & Blandon, 1982; Salzman & Munhall, 1989; Keating, 1990; Farnetani & Recasens, 1999), uno dei più convincenti sembra essere il modello di coarticolazione proposto da Cohen e Massaro (Cohen & Massaro, 1993) e basato sulla *gestural theory of speech production* di Löfqvist (Löfqvist, 1990; Munhall & Löfqvist, 1992). Secondo la teoria di Löfqvist, ad ogni singolo gesto articolatorio è associata una funzione di ***dominanza*** con le stesse caratteristiche dei gesti fonetici. La funzione di *dominanza* ha lo scopo di prevedere l'estensione anticipatoria

(all'indietro) e l'estensione perseverativa (in avanti) di un segmento rispetto all'altro, ed è caratterizzata da una propria ampiezza, durata, e grado di attivazione. L'ampiezza determina l'importanza relativa del gesto per il segmento relativo; 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. La *dominanza D* è una funzione esponenziale asimmetrica del tipo:

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

in cui $\tau$ rappresenta la distanza temporale dal centro del segmento fonetico, $\alpha$ rappresenta l'ampiezza della *dominanza*, $\theta$ l'estensione all'indietro (*bw*) o in avanti (*fw*) dell'influenza del segmento, e l'esponente c determina il grado di attivazione o pendenza della curva relativa, e può anche essere interpretato come grado di rilascio del movimento articolatorio.

Il metodo implementato da Cohen e Massaro è stato migliorato per realizzare transizioni più accurate fra *target* articolatori successivi e per risolvere parecchie difficoltà incontrate nella modellizzazione articolatoria delle consonanti bilabiali e labiodentali. Questo obiettivo è stato raggiunto in due modi:

- Adottando una nuova versione più generale delle funzioni di *dominanza*, che utilizza effettivamente l'esponente c dell'equazione (1) per adattarsi alla diversa velocità del parlato.
- Aggiungendo al modello originale due nuove componenti dette ***resistenza temporale*** e ***forma***.

La funzione ***forma S*** ha l'effetto di modellare l'andamento del target articolatorio in prossimità del suo massimo rilievo, per cui otterremo un target non più discreto, ma variante nel tempo con una propria caratteristica. La funzione *forma* risulta utile nel riprodurre andamenti con caratteristiche particolari come ad esempio la pendenza rilevata nella produzione della vocale /u/ in alcuni contesti consonantici. 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.

L'equazione per questa funzione è:

$$S_{LA}(\tau) = \begin{cases} \beta_{bw} \left| \frac{\tau}{h_{bw}} \right|^{p_{bw}} + 1 \text{ se } \tau < 0 \\ \beta_{fw} \left| \frac{\tau}{h_{fw}} \right|^{p_{fw}} + 1 \text{ se } \tau > 0 \end{cases} \quad (2)$$

dove *LA* indica una *forma* di tipo *Look-Ahead*, in cui l'influenza della funzione è proporzionale alla distanza con il target successivo o antecedente; dove $h_{bw}$ e $h_{fw}$ rappresentano fattori proporzionali alla distanza dai target precedenti o successivi;

La funzione di ***resistenza temporale R*** è stata introdotta per 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, 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.

$$R(\tau) = \begin{cases} e^{-6\left|\frac{\tau}{h_{bw}}\right|^4} & \text{se } \tau < 0 \\ e^{-6\left|\frac{\tau}{h_{fw}}\right|^4} & \text{se } \tau > 0 \end{cases} \qquad (3)$$

dove $h_{bw}$ e $h_{fw}$ rappresentano come nella funzione *forma* fattori proporzionali alla distanza dai target precedenti o successivi

La formula completa della nuova funzione è la seguente:

$$F_{new}(t) = \frac{\sum_{i=1}^{N} T_i \cdot S_i(t-t_i) \cdot R_i(t-t_i) \cdot D_i(t-t_i)}{\sum_{i=1}^{N} R_i(t-t_i) \cdot D_i(t-t_i)} \qquad (4)$$

dove $N$ è il numero dei fonemi interessati, $D$ è la *dominanza* relativa all'i-esimo segmento calcolata secondo l'equazione (1), $S$ è la funzione *forma* data dalla formula (2) e $R$ la *resistenza temporale* derivata dall'equazione (3).

La procedura di stima dei parametri è basata su metodo classico di minimizzazione dei minimi quadrati:

$$e = \sum_{r=1}^{R}\left(\sum_{n=1}^{N}\left(Y_r(n) - F_r(n)\right)^2\right) \qquad (5)$$

fra i dati reali *Y(n)* e le curve ottenute in uscita dal modello modificato *F(n)*, rappresentato dall'equazione (4), su un certo numero di ripetizioni $R$ dello stesso tipo di sequenze (Fig. 22).

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, la funzione costo presenta molti minimi globali e deve essere necessariamente guidata in modo manuale verso gli opportuni valori di target finali. Particolare attenzione è stata rivolta alla selezione del metodo di ottimizzazione. Essendo il numero di parametri in gioco molto alto (Fig. 21), 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.

In Fig. 21 si possono vedere i coefficienti ottimizzati relativi alle varie funzioni *dominanza*, *forma* e *resistenza temporale* per la transizione /aba/. La presenza, o l'assenza, del segno di spunta vicino al valore indica che quel parametro è in quel momento soggetto al processo di ottimizzazione, oppure tenuto costante. In basso compaiono i valori di ampiezza e la posizione temporale dei *target* fonetici reali misurati dai dati di acquisizione.

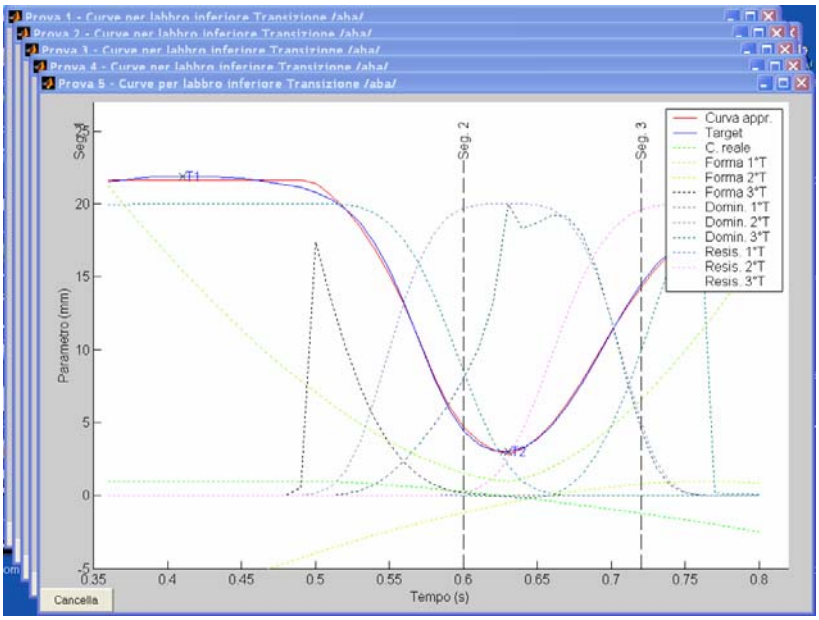


Figura 22 – Optimize: Curva reale (in blu) e approssimata (in rosso) della transizione /aba/.

In Fig. 22 si può vedere la curva temporale del labbro inferiore nella pronuncia di /aba/ e la curva approssimata secondo l'equazione (4) con i valori impostati di Fig. 21. Compaiono anche le linee tratteggiate dei contributi separati delle relative funzioni di *dominanza*, *forma* e *resistenza temporale.*

## 11. ANIMAZIONE DA XML E DA TESTO SCRITTO

Le Facce Parlanti hanno una applicazione ovvia nella lettura da testo scritto (***Text-to-Animation Synthesis***). Il programma che prepara il flusso di comando audio-visuale per un agente di animazione, è stato chiamato **AVengine**. Può ricevere in ingresso tanto un puro testo scritto, quanto un testo contenente parole chiave, come ad esempio un testo **XML**.

Nel caso del testo semplice, la sintesi audio e l'animazione video sono determinate unicamente dalla sequenza dei fonemi del testo in ingresso. Nel caso dell'audio, il risultato dipende anche dal tipo di *database* fonetico, che si stia utilizzando, e dalle regole prosodiche che eventualmente si siano implementate (Tesser *et alii*, 2003,2004). In questo caso, come si diceva nell'introduzione, il risultato sonoro non può cambiare in relazione al contenuto del messaggio sintetizzato.

Per aggiungere espressività e emozioni alle *Talking Heads*, è necessario ricorrere a opportuni attributi, o *tag*, che siano poi interpretati dal motore di animazione, o meglio ad opportuni linguaggi che ne definiscano la grammatica, come può essere appunto l'**XML**.

Come è noto, il linguaggio **XML** (***Extensible Markup Language***) è stato creato nel 1996 dal W3C, e cioè il *World Wide Web Consortium*, come una specie di dialetto del più generale **SGML** (***Standard Generalized Markup Language***) che è lo standard internazionale di comunicazione sulla rete World Wide Web. La principale caratteristica di SGML e XML è che non sono semplicemente dei linguaggi di *markup* come l'**HTML** (***HyperText Markup Language***), ma meta-linguaggi che danno la possibilità all'utilizzatore stesso di definire la struttura del linguaggio, e cioè le parole chiave (o sintassi) e le regole (o grammatica), che descrivono le relazioni fra la struttura e il contenuto del documento. Questo insieme di definizioni è detto ***Document Type Definition*** (**DTD**).

Altre caratteristiche che rendono **XML,** un linguaggio potente e flessibile, sono:

- L'estensibilità, che non pone limiti alla complessità lessicale e sintattica del linguaggio (a differenza dell'HTML, che ha un numero fisso di attributi).
- La possibilità di verificare la correttezza del testo XML in base alle definizioni date nel DTD.
- L'interoperabilità, che permette di condividere e riutilizzare il documenti sulla rete fra applicazioni e piattaforme hardware diverse.
- Il fatto che XML è un *software Open Source*.

Per permetter l'animazione espressiva delle *Talking Heads*, è stato creato un apposito DTD, chiamato **APML**, ***Affective Presentation Markup Language*** (De Carolis *et alii*, 2004). Il documento è un tentativo di definizione del comportamento degli agenti conversazionali, con l'introduzione di *tag* tipici degli atteggiamenti assunti nel corso di un dialogo, come ad esempio*: implore, order, suggest, propose, warn, approve, praise, recognize, disagree*, ecc., che fanno parte dell'elemento *performative*.

Si possono isolare quattro gruppi di funzioni comunicative:

- Le convinzioni dell'agente virtuale (attributo *certainty*, ecc.).
- Le sue intenzioni (attributi *performative, comment, belief-relation, turn-allocation*, ecc.).
- Il suo stato affettivo (attributo *affective*)
- Lo stato metacognitivo mentale (*i'm thinking, i'm planing*, ecc.)

In questo filmato si può vedere un esempio di sintesi relativa: 

È stata sviluppata una versione estesa del linguaggio APML, per aggiungere alla sintesi espressiva delle Facce Parlanti anche una sintesi audio, che simuli i correlati acustici caratteristici delle differenti emozioni (Drioli *et alii*, 2003, 2004). In questo modo i *tag* del linguaggio APML sono utilizzati per controllare tanto la parte visiva che la parte audio necessarie all'animazione.

Il nuovo modulo APML/VSML (dove VSML sta per *Voice Signal Markup Language*) ha una architettura gerarchica a tre livelli (Fig. 23). A livello più elevato di astrazione troviamo i *tag* relativi alle emozioni: *<anger>, <joy>, <fear>, <sadness>, <surprise>,* ecc. Questi ultimi sono a loro volta descritti a livello intermedio con la terminologia tipica della ***Voice Quality***, in termini cioè di modalità di fonazione: *<modal>, <soft>, <pressed>, <breathy>, <whispery>, <creaky>*, ecc. Questi *tag* sono, infine, definiti al livello più basso della struttura in termini di caratteristiche acustiche basilari come: *<spectral tilt>, <shimmer>, <jitter>*, ecc.

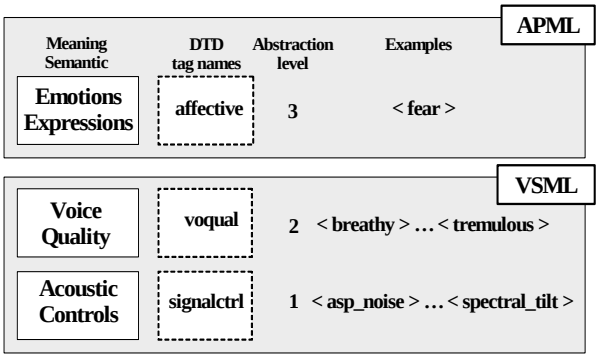


Figure 23 – Struttura del linguaggio APML/VSML per la sintesi audio delle emozioni.

La mappatura degli attributi APML avviene dall'alto al basso, come nell'esempio di Fig. 23, in cui la paura <fear>, è interpretata con una tipologia di fonazione soffiata e tremolante (*<breathy>, <tremolous>*), ed espressa infine in termini di parametri fisici, come rumore di aspirazione, bilancio spettrale alte-basse frequenze, e modulazione della F0 (*<asp. noise>, <spectral tilt>, <F0 modulation>,* ecc.).

L'implementazione del nuovo APML/VSML ha richiesto una modifica tanto del *parser* sintattico di FESTIVAL (Black *et alii*, FESTIVAL Home Page), che del sintetizzatore audio **Mbrola** (Mbrola Home Page), per poter operare le trasformazione opportune nel tempo ed in frequenza del *pitch*, dell'ampiezza e del contenuto spettrale dei segmenti in questione. Una caratteristica interessante dell'implementazione è che l'applicazione delle trasformazione non è statica, ma è resa dinamica mediante l'uso di generatori di inviluppo opportunamente comandati.

Come si era detto nell'introduzione, oltre al sistema Mbrola, in questa versione di InterFace è stata implementata la tecnica SMS (Spectral Modeling Synthesis) (SMS Home Page).

Rispetto a Mbrola, che lavora nel dominio temporale, giustapponendo forme d'onda preesistenti, SMS opera nel dominio della frequenza, sommando le diverse parziali sinusoidali del suono (parte deterministica) e assieme ad una componente di rumore (parte residuale stocastica). Nel caso di una generica sintesi vocale, SMS è computazionalmente più dispendiosa e qualitativamente meno efficace di Mbrola, dovendo generare decine di sinusoidi per ogni periodo di ogni segmento. Ma, se non ci si accontenta di una qualità vocale immutabile e si vogliono ottenere trasformazioni timbriche come avvengono nel parlato emotivo, SMS offre un vantaggio importante. In questo caso, l'utilizzo di Mbrola è più problematico, perché occorre convertire il segmento di forma d'onda nella sua

rappresentazione spettrale, per operare le dovute trasformazioni (inviluppo spettrale, *time stretching*, *pitch shifting*, ecc.) senza alterare le caratteristiche formantiche originali. Nella successiva riconversione a forma d'onda, ci si scontra con il problema delle discontinuità di fase che devono essere controllate per la concatenazione dei difoni.

Un diverso approccio alla simulazione delle emozioni può essere realizzato partendo dall'informazione prosodica (*pitch* e durata) contenuta in un parlato reale. La metodologia seguita all'ISTC è stata quella di creare un *database* apposito (Emotional-CARINI, E-Carini), facendo leggere ad un attore professionista una novella (*Il Colombre* di Dino Buzzati, secondo le 6 emozioni fondamentali, oltre alla lettura neutrale. Complessivamente ogni emozione ha una durata di 15 m.

Le caratteristiche prosodiche dei racconti sono state modellate con CARTs (*Classification And Regression Trees*) (Breiman *et alii*, 1984, Cosi *et alii*, 2002d, Tesser, 2004).

Un esempio di animazione con l'audio generato in questo modo può essere visto qui: 

## 12. ANIMAZIONE DA *FILE* AUDIO

Un problema, che si incontra quando si voglia applicare una voce reale ad un agente virtuale, è ovviamente quello di sincronizzare il flusso visuale con quello sonoro preesistente. Dentro InterFace, si è voluto creare uno strumento che permetta di segmentare automaticamente un *file* audio, e di ricavare la sequenza fonetica con le relative durate, necessaria per pilotare l'animazione sincrona con il parlato di partenza.

L'allineamento audio-visuale è ottenuto con un riconoscitore della voce dell' ***OGI Toolkit*** basato su una rete neurale con architettura ibrida HMM/ANN (OGI - CSLU toolkit - OGI School of Science & Engineering - Center for Spoken Languages Understanding).

In alternative, in questa versione di InterFace, è stato integrato il riconoscitore ***Sonic*** http://cslr.colorado.edu/beginweb/speech_recognition/sonic.html (Pellom, 2003), che offre prestazioni superiori in termini di tempo di esecuzione e *Word Error Rate* (WER).

Per quanto riguarda l'italiano, entrambi i riconoscitori sono stati allenati sull'*Acoustic-Phonetic and Spontaneous Speech Corpus* (APASCI) dell'IRST (*Istituto per la Ricerca Scientifica e Tecnologica – Trento*).

Come si era anticipato nell'introduzione, si realizza per questa via una tipica ***Wav-to-Animation Synthesis***.

Nell'esempio di animazione che segue, l'audio, sintetizzato con Loquendo, è stato sincronizzato con l'animazione nel modo sopra descritto 

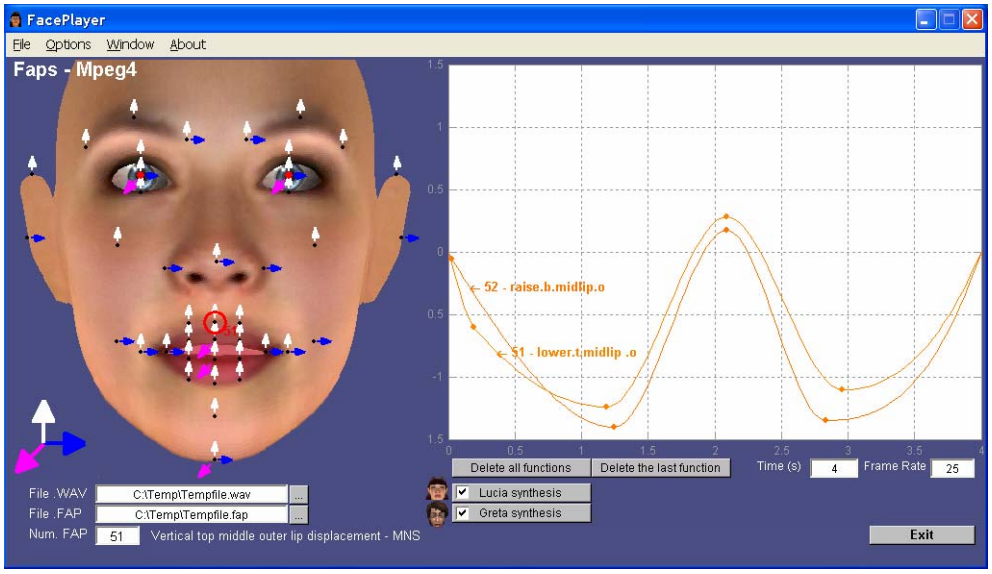


Figura 24 – FacePlayer: Finestra principale del programma.

## 13. FACEPLAYER E EMOTIONPLAYER

Per lo sviluppo efficace delle applicazioni delle *Talking Heads*, è importante disporre di mezzi che permettano il test rapido dei parametri relativi all'animazione audio-visuale.

**FacePlayer** permette di controllare il movimento di un unico FAP (o di un insieme voluto di FAPs), mediante funzioni schematizzate per punti ed poi interpolate. In Fig. 24, si può vedere la finestra principale del programma. Sulla faccia di sinistra le frecce indicano le possibili componenti del movimento di alterazione della superficie della pelle: bianco per i movimenti verticali, blu per quelli orizzontali, magenta per la direzione antero-posteriore. Il punto da controllare con la funzione voluta è scelta con un clic su una delle frecce, mentre nel riquadro a destra della figura si traccia la curva con una serie di punti. Anche in questo caso, come nella scelta della configurazione *marker*-FAP (cap. 5.4), si possono alterare i valori di scala, ovverosia i *Facial Animation Parameter Units* (FAPU), per adattare il *FAP-stream* alle dimensioni di una particolare faccia.

In Fig. 25 è riportato una sequenza di animazione con il movimento del labbro inferiore e superiore dedotto dalle funzioni disegnate in Fig. 24.

Figura 25 – EmotionPlayer: Sequenza (parziale) di controllo dell'apertura verticale delle labbra corrispondente alle funzioni definite in Fig. 24, con compressione delle labbra al fotogramma 5.

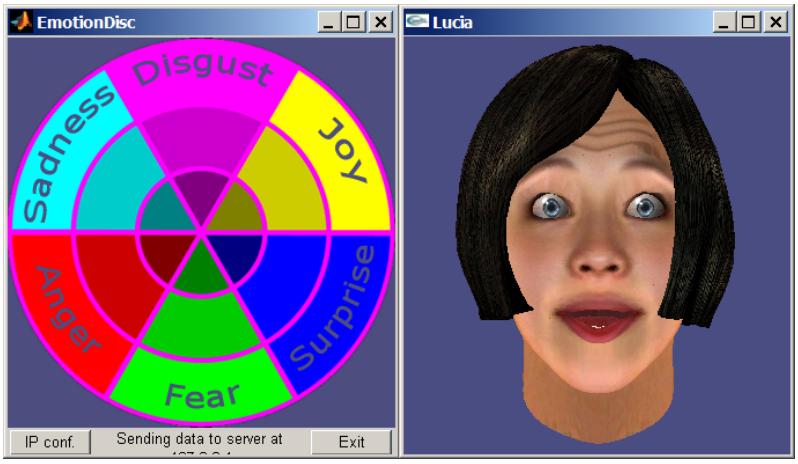


Figura 26 – EmotionPlayer: Premendo sulle differenti zone del disco a sinistra si ottiene la sintesi della corrispondente emozione.

**EmotionPlayer** è derivato dal lavoro di Zofia Ruttkay (Ruttkay *et alii*, 2003) per consentire la sintesi prototipale delle emozioni principali (Fig. 26).

Il programma presenta un disco diviso in settori radiali corrispondenti alle sei principali emozioni (gioia, rabbia, paura, sorpresa, tristezza, disgusto), e in tre zone concentriche, corrispondenti a tre diversi livelli di intensità dell'emozione relativa. Premendo su una delle diverse zone si ottiene un set di funzioni di controllo dei FAPs, predisposte e modificabili, le cui ampiezze dipendono dal livello di intensità emotiva scelto, rispettivamente basso, medio, ed elevato (Fig. 27).

Nella Fig. 28 in basso, compare una sequenza di espressioni emotive reali dovute all'attore Fabio Fusco, e in alto una sequenza virtuale corrispondente, ottenuta con di EmotionPlayer.

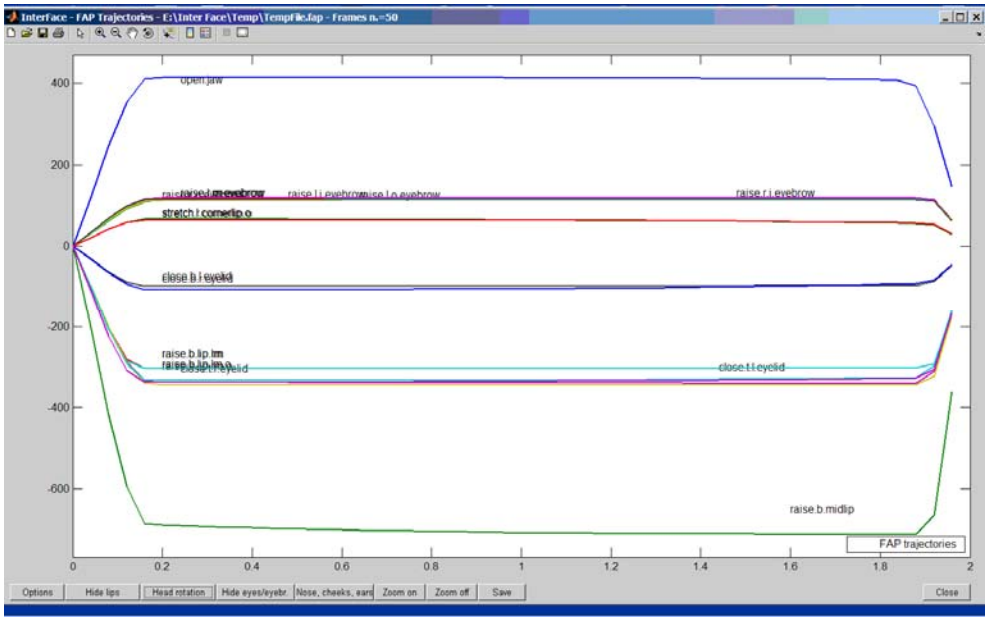


Figura 27 – EmotionPlayer: Funzioni di controllo per la "*Sorpresa*"

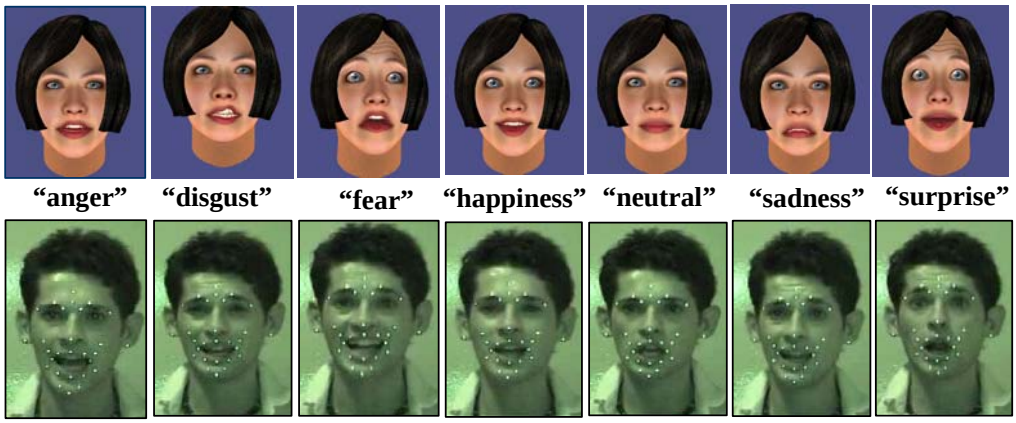


Figura 28 – EmotionPlayer: Simulazione prototipale delle principali emozioni. In basso una sequenza reale interpretata dall'attore Fabio Fusco.

## 14. RINGRAZIAMENTI

Il lavoro è stato in parte finanziato dal progetto PF-STAR (*Preparing Future multiSensorial inTerAction Research*, European Project IST-2001-37599, pfstar.itc.it), e TICCA (*Tecnologie cognitive per l'Interazione e la Cooperazione Con Agenti artificiali*, in cooperazione fra il CNR e la Provincia Autonoma Trentina).

InterFace è stato realizzato con il prezioso contributo di: Carlo Drioli, Vincenzo Ferrari, Andrea Fusaro, Daniele Grigoletto, Enrico Marchetto, Giulio Paci, Giulio Perin, Fabio Rossi (B|T|S) e Fabio Tesser.

## 15. BIBLIOGRAFIA


Adjoudani, C. Benoit. (1995), Audio-Visual Speech Recognition Compared Across Two Architectures, Proc. Eurospeech-95, Madrid (Spain), Vol. 2., 1563-1566.

Al-Bamerni A., Blandon A. (1982), One-stage and two-stage patterns of velar coarticulation, *Journal of the Acoustical Society of America*, Vol. 72, Suppl. 1, 104.

ANVIL Home Page: http://www.dfki.de/~kipp/anvil/documentation.html

B|T|S Home Page: www.bts.it

Banse, R., Scherer, K. R. (1996) Acoustic profiles in vocal emotion expression, *Journal of Personality and Social Psycology*, 70, 614-636.

Bell-Berti F., Harris K.S. (1981), A Temporal Model of Speech Production, *Phonetica*, 1981, Vol. 38, 9-20.

Benoit C., Lallouache T., Mohamadi T., Abry C. (1992), A Set of French Visemes for Visual Speech Synthesis, in Bailly G., Benoit C., Sawallis T.R. (Eds.), *Talking machines: Theories, Models, and Designs*, North-Holland, Amsterdam, 485-504.

Benoit C., Guiard-Marigny T., Le Goff B., Adjoudani A. (1996), Which Components of the Face Do Humans and Machines Best Speech-Read?, in Stork D. and Hennecke M. (Eds.) *Speech-reading by humans and machine: models, systems and applications*, Springer-Verlag, New York, 315-328.

Beskow J. (1995), Rule-Based Visual Speech Synthesis, in *Proceedings of Eurospeech '95, 4th European Conference on Speech Communication and Technology*, Madrid, 299-302.

Bilvi M.(2002), *Progetto e Sviluppo di un Agente Conversazionale Multimodale: Animazione e Sincronizzazione dei Segnali Verbali e non Verbali*, M.S. Thesis, La Sapienza University, Rome.

Biscetti S., Cosi P., Delmonte R., Cole R. (2004), Italian Literacy Tutor: un adattamento all'italiano del 'Colorado Literacy Tutor'", in *Atti DIDAMATICA 2004*, Ferrara (Italy), 249-253.

Black A., Taylor P., Caley R., Clark R., *The Festival Speech Synthesizer*, www.cstr.ed.ac.uk/projects/festival

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

Breiman L., Friedman J., Olshen R., and Stone C., *Classification and regression trees*. Wadsworth and Brooks, 1984.

Cahn J. (1990), The Generation of Affect in Synthesized Speech, *Journal of the American Voice I/O Society*, Vol. 8, 1-19.

Chen T., Rao R. (1998), Audio-Visual Integration in Multimodal Communications, *Proc. of the IEEE*, Vol. 86, no. 5, 837-852.

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

Cohen M., Massaro D. (1990), Synthesis of Visible Speech, *Behaviour Research Methods, Instruments and Computers*, Vol. 22 (2), 260-263.

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

Cohen M., Walker R., Massaro D. (1996), Perception of Synthetic Visual Speech, in Stork D. and Hennecke M. (Eds.), *Speech-reading by Humans and Machine: Models, Systems and Applications*, Springer-Verlag, New York, 153-168.

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

Cosi P., Magno C. E. (1996), Lip and Jaw Movements for Vowels and Consonants: Spatio-Temporal Characteristics and Bimodal Recognition Applications, in Stork D. and Hennecke M. (Eds.) *Speech-reading by Humans And Machine: Models, Systems and Applications*, Springer-Verlag, New York, 291-313.

Cosi P., Tesser F., Gretter R., Avesani, C. (2001), Festival Speaks Italian!, *Proc. Eurospeech 2001*, Aalborg, Denmark, 509-512.

Cosi P., Magno C. E., Perin G., Zmarich C. (2002a), Labial Coarticulation Modeling for Realistic Facial Animation, *Proc. ICMI 2002*, *4th IEEE International Conference on Multimodal Interfaces 2002*, Pittsburgh (USA), 505-510.

Cosi P., Magno C. E., Tisato G., Zmarich C. (2002b), Biometric Data Collection for Bimodal Applications, *Proc. of COST 275 Workshop*, *The advent of Biometric on the Internet*, Rome, 127-130.

Cosi P., Ferrari V., Magno C. E., Perin G., Tisato G., Zmarich C. (2002c), GRETA e LUCIA: Due Realistiche Facce Parlanti Animate Mediante un Nuovo Modello di Coarticolazione, in *Atti delle XIII Giornate di Studio GFS 2002*, Pisa, 27-134.

Cosi P., Avesani C., Tesser F., Gretter R., and Pianesi F. (2002d), On the use of Cart-Tree for prosodic predictions in the Italian Festival TTS, in Cosi P. Magno E. Zamboni A. editors, *Voce, Canto, Parlato- Studi in onore di Franco Ferrero*, UNIPRESS Padova, Italy, 73–81.

Cosi P., Fusaro A., Tisato G. (2003a), LUCIA a New Italian Talking-Head Based on a Modified Cohen-Massaro's Labial Coarticulation Model, *Proc. Eurospeech 2003*, Vol. III, 2269-2272.

Cosi P., Magno C. E. (2003b), *E-learning* e Facce Parlanti: nuove applicazioni e prospettive, in *Proc. of XIV Giornate di Studio del G.F.S.*, Viterbo, Italy, (in press).

Cosi P., Delmonte R., Biscetti S., Cole R. A., Pellom B., van Vuren S. (2004a), Italian Literacy Tutor, tools and technologies for individuals with cognitive disabilities, *Proc. of InSTIL/ICALL Symposium 2004*, Venice (Italy), 207-215.

Cosi P., Fusaro A., Grigoletto D., Tisato G. (2004b), Data-Driven Tools for Designing Talking Heads Exploiting Emotional Attitudes, in *Proc. of Tutorial and Research Workshop, Affective Dialogue Systems*, Kloster Irsee (Germany), 101-112.

Cosi P., Drioli C., Tesser F., Tisato G. (2005), INTERFACE toolkit: a new tool for building IVAs, in *Proceedings of IVA (Intelligent Virtual Agents) 2005*, Kos, Greece, September 2005, pp. 75-87.

Cowie R., Douglas-Cowie E., Tsapatsoulis N., Votsis G., Kollias S., Fellenz W., Taylor J. (2001), Emotion Recognition in Human-Computer Interaction, *IEEE Signal Processing Magazine*, Vol. 8, no. 1, 32-80.

D'Amico M., Ferrigno G. (1990), Technique for the Evaluation of Derivatives from Noisy Biomechanical Displacement Data Using a Model-Based Bandwidth-Selection Procedure, *Medical & Biological Engineering & Computing*, 28, 407-415.

D'Amico M., Ferrigno G. (1992), Comparison Between the More Recent Techniques for Smoothing and Derivative Assessment in Biomechanics, *Medical & Biological Engineering & Computing*, 30, 193-204.

Damper R. (2001), Learning About Speech from Data: Beyond NETtalk, in *Data-Driven Techniques in Speech Synthesis*, R.I.Damper Ed., Kluwer Academic Publisher, 1-25.

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

De Carolis B., Pelachaud C., Poggi I., Steedman M. (2004), APML, a Markup Language for Believable Behavior Generation, in Prendinger H. & Ishizuka M. (Eds.), *Life-like Characters. Tools, Affective Functions and Applications*. Springer-Verlag, Berlin, 65-86.

Doenges P., Capin T., Lavagetto F., Ostermann J., Pandzic I., Petajan E. (1997), MPEG-4: Audio/video and Synthetic Graphics/Audio for Real-Time, Interactive Media Delivery, *Signal Processing: Image Communication*, 9(4), 433-463.

Douglas_Cowie E., Campbell N. (Eds.) (2003), *Special Issue on Speech and Emotion*, in *Speech Comm.*, 40 (1-2), 1-258.

Drioli C., Tisato G., Cosi P., F. Tesser. (2003), Emotions and Voice Quality: Experiments with Sinusoidal Modeling, *Proc. of Voqual 2003, Voice Quality: Functions, Analysis and Synthesis, ISCA Workshop*, Geneva (Switzerland), 127-132.

Drioli C., Tisato G., Cosi P., F. Tesser. (2004), Control of Voice Quality for Emotional Speech Synthesis, *CD-Rom Proceedings of AISV 2004*, EDK Editore s.r.l., Padova, 2005, 789-798.

Ekman P., Friesen W. (1977), *Manual for the Facial Action Coding Systems*, Consulting Psychologist Press Inc., Palo Alto (USA).

Ekman P., Friesen W. (1978), *Facial Action Coding System*, Consulting Psychologist Press Inc., Palo Alto (USA).

Farnetani E., Recasens D. (1999), CoarticulationModels in Recent Speech Production Theories, in Hardcastle W.J. (Eds.), *Coarticulationin Speech Production*, Cambridge University Press, Cambridge, 31-68.

Ferrigno G., Pedotti A. (1985), ELITE - A digital dedicated hardware system for movement analysis via real-time TV signal processing. *IEEE Transactions on Biomedical Engineering*, vol. 32., 943-950.

Festival Home Page: www.cstr.ed.ac.uk/projects/festival/

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

Iida A., Campbell N., Higuchi F., Yasumara M. (2003), A Corpus-Based Speech Synthesis System with Emotion, *Speech Comm.*, Vol. 40, 161-187.

InterFace Home Page: www.pd.istc.cnr.it/interface

Keating P. (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*, 451–470. Cambridge University Press.

Keltner D., Ekman P., Gonzaga G., Beer J. (2003), Facial Expression of Emotions, in R. Davidson, H. Goldsmith, K. Scherer (Eds.) *Handbook of the Affective Sciences*, Oxford University Press, New York, 415-432.

Michael Kipp (2004), Gesture Generation by Imitation - From Human Behavior to Computer Character Animation, Boca Raton, Florida: Dissertation.com

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

Lavagetto F., Pockaj R. (1999), The Facial Animation Engine: Towards a High-Level Interface for the Design of Mpeg-4 Compliant Animated Faces, *IEEE Trans. on Circuits and Systems for Video Technology*, 9(2), 277-289.

Le Goff B., Guiard-Marigny T., Cohen M., Benoît C. (1994), Real-time Analysis-Synthesis and Intelligibility of Talking Faces, in *Proc. of the second ESCA/IEEE Workshop on Speech Synthesis*, New Paltz, New York, 53-56.

Le Goff B., Benoît C. (1996), A Text-To-Audiovisual-Speech Synthesizer for French, in *Proc. of ICSLP '96: The Fourth International Conference on Spoken LanguageProcessing*, Philadelphia, 2163-2166.

Le Goff B. (1997), *Synthèse A Partir du Texte de Visages 3D Parlant Français*, Ph.D. Thesis, Grenoble, France.

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

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

Loquendo Home Page: www.loquendo.it

Magno C. E., Vagges K., Ferrigno G., Zmarich C. (1993), Articulatory Dynamics of Lips in Italian /'VpV/ and /'VbV/ sequences *Proc. of Eurospeech'93*, Berlin, Vol. 1, 409-413.

Magno C. E., Zmarich C., Cosi P., Ferrero F. (1997), Italian Consonantal Visemes: Relationships Between Spatial/Temporal Articulatory Characteristics and Co-Produced Acoustic Signal, in C. Benoit and R. Campbell, (Eds.), *Proc. of AVSP'97*, Rhodes (Greece), 5-8.

Magno C. E., Zmarich C., Cosi P. (1998), Statistical Definition of Visual Information for Italian Vowels and Consonants, in D. Burnham, J. Robert-Ribes, E. Vatikiotis-Bateson, (Eds.), *Proc. of AVSP'98*, Terrigal-Sydney (Australia), 135-140.

Magno C. E., Poggi I. (2001), Dall'Analisi della Multimodalità Quotidiana alla Costruzione di Agenti Animati con Facce Parlanti ed Espressive, in Magno C. E. and Cosi P. (Eds.) *Multimodalità e Multimedialità della Comunicazione, Atti delle XI Giornate di Studio del GFS*, Unipress, Padova, 47-55.

Magno C. E., Cosi P., Drioli C., Tisato G., Cavicchio F. (2003), Co-production of Speech and Emotion: Bi-Modal Audio-Visual Changes of Consonant and Vowel Labial Targets, *Proc. AVSP 2003, Audio Visual Speech Processing, ISCA Workshop*, St. Jorioz (France), 209-214.

Magno C. E., Cosi P., Drioli C., Tisato G., Cavicchio F. (2004), Modifications of Phonetic Labial Targets in Emotive Speech: Effects of the Co-Production of Speech and Emotions, *Speech Communication: Special issue on audio visual speech processing*, Vol. 44, 173-185.

Massaro D. (1987), Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry, in Dodd B. and Campbell R. (Eds.), *Hearing by Eye: The Psychology of Lip-Reading*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 53-83.

Massaro D. (1996), Bimodal Speech Perception: A Progress Report, in Stork D. and Hennecke M. (Eds.) *Speech-reading by Humans And Machine: Models, Systems and Applications*, Springer-Verlag, New York, 79-102.

Massaro D. (1998), *Perceiving Talking Faces: from Speech Perception to a Behavioral Principle*, A Bradford book, the MIT Press, Cambridge, Massachusetts.

Massaro, D., and Egan, P. (1998), Perceiving Affect from the Voice and the Face, *Psychological Bulletin*, 3, 1021-1032.

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

Mbrola Home Page: www.tcts.fpms.ac.be/synthesis/mbrola

McGurk H., MacDonald J. (1976), Hearing Lips and Seeing Voices, *Nature*, 264, 746-748.

MPEG-4 Standard Home Page: http://mpeg.telecomitalialab.com/standards/MPEG-4

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

OGI toolkit [OGI - CSLU toolkit - OGI School of Science & Engineering - Center for Spoken Languages Understanding].

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

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

Parke F. (1974) *A Parametrical Model for Human Face*, Ph.D. Thesis, *Tech. Report UTEC-CSc-75-047*, University of Utah, Salt Lake City (USA).

Parke F. (1982), Parametrized Models for Facial Animation, *IEEE Computer Graphics*, 2(9), 61-68.

Pasquariello S. (2000), *Modello per l'Animazione Facciale in MPEG-4*, M.S. Thesis, University of Rome.

Pearce, B. Wyvill, G. Wyvill, D. Hill (1986), Speech and Expression: A Computer Solution to Face Animation, *Graphic and Vision '86*, 136-140.

Pelachaud C., Magno C. E., Zmarich C., Cosi P. (2001), Modeling an Italian Talking Head, *Proc. AVSP 2001*, Aalborg, Denmark, 72-77.

Pellom B., Hacioglu K., "Recent Improvements in the CU Sonic ASR System for Noisy Speech: The SPINE Task", *ICASSP*, Hong Kong, vol. I, 2003, pp. 4-7.

Perin G. (2001), *Facce Parlanti: Sviluppo di un Modello Coarticolatorio Labiale per un Sistema di Sintesi Bimodale*, M.S. Thesis, Univ. of Padova.

Petajan E. (1984), *Automatic Lip-Reading to Enhance Speech Recognition*, Ph.D. Thesis, Univ. of Illinois at Urbana-Champaign.

Prendinger H. & Ishizuka M. (Eds.) (2004), *Life-like Characters. Tools, Affective Functions and Applications*. Springer-Verlag, Berlin, 65-86.

Rank E., Pirker H.(1998), Generating Emotional Speech with a Concatenative Synthesizer, in *Proc. 5th International Conference on Spoken Language Processing (ICSLP'98),* Sydney, Australia.

Ruttkay Z., Noot H., ten Hagen P. (2003), Emotion Disc and Emotion Squares: Tools to Explore the Facial Expression Space, *Computer Graphics Forum*, 22(1), 49-53.

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

Scherer K. (2003), Vocal Communication of Emotion: A Review of Research Paradigm, *Speech Comm.*, 40, 227-256.

Scherer K., Johnstone T., Klasmeyer G. (2003), Vocal Expression of Emotion. In R. J. Davidson, H. Goldsmith, K. R. Scherer (Eds.), *Handbook of the Affective Sciences*, 433-456. New York, Oxford University Press.

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

Serra, X. and Smith, J. O. (1990), Spectral Modeling Synthesis: A Sound Analysis/Synthesis Based on a Deterministic plus Stochastic Decomposition", *Computer Music Journal*, vol. 14(4), 1990.

Silsbee P., Allen A.C. (1993), Medium-Vocabulary Audio-Visual Speech Recognition, Proc. NATO ASI, *New advances and trends in speech recognition and coding*, 13-16.

SMS Home Page: http://www.iua.upf.es/~sms

Soderkvist I. and Wedin P. (1993), Determining the movements of the skeleton using well-configured markers, *Journal of Biomechanics*, 26:1473-1477.

Sommavilla, G., Drioli, C., Cosi, P. (2005), "Sintesi vocale concatenativa per l'italiano tramite modello sinusoidale", *CD-Rom Proceedings of AISV 2005*, EDK Editore s.r.l., Padova, 2005, 761-772.

Sonic Home Page: http://cslr.colorado.edu/beginweb/speech_recognition/sonic.html

Stork D., Henneke M. (Eds.) (1996), *Speech-Reading by Humans and Machine: Models, Systems and Applications*, Springer-Verlag, New York.

Stork D., Wolff G., Levine E. (1992), Neural Network Lip-Reading System for Improved Speech Recognition, *Proc. of IEEE International Joint Conference on Neural Networks*, *IEEE-IJCNN-92*, 285-295.

Summerfield Q. (1987), Some Preliminaries to a Comprehensive Account of Audio-Visual Speech Perception, in Dodd B. and Campbell R. (Eds.), *Hearing by Eye: The Psychology of Lip-Reading*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 3-51.

Tesser F., Cosi P., Drioli C., Tisato G., (2004), Prosodic Data-Driven Modelling of a Narrative Style in FESTIVAL TTS, in *Proc. of the 5th ISCA Speech Synthesis Workshop*, Pittsburgh (USA), 185-190.

Tesser F., Cosi P., Drioli C., Tisato G., (2004), Modelli Prosodici Emotivi per la Sintesi dell'italiano, *CD-Rom Proceedings of AISV 2004*, EDK Editore s.r.l., Padova, 2005.

Tesser F., Cosi P., Drioli C., and Tisato G. (2004), Prosodic data driver modelling of a narrative style in Festival TTS", *CDROM Proc. of 5th ISCA Speech Synthesis Workshop*, Pittsburgh, PA, U.S.A.

Tisato G., Cosi P., Drioli C., Tesser F. (2004a), InterFace: Strumenti Interattivi per l'Animazione delle Facce Parlanti, *CD-Rom Proceedings of AISV 2004*, EDK Editore s.r.l., Padova, 2005, pp. 817-846.

Tisato G., Cosi P., Drioli C., Tesser F.(2005b), "InterFace: New Tool for Building Emotive/Expressive Talking Heads", *Proceedings of INTERSPEECH 2005*, Lisbon, Portugal, 2005, pp. 781-784.

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

Walden B., Prosek R., Montgomery A., Scherr C., Jones C. (1977), Effects of Training on the Visual Recognition of Consonants, *Journal of Speech and Hearing Research*, Vol. 20, 130-145.

Wolf R. (1983), *Elements of Photogrammetry*, Mc Graw-Hill Publisher.

Zierdt, A. (1993), Problems of Electromagnetic Position Transduction for a Three-dimensional Articulographic Measurement System, *FIPKM* 31, 1993, pp. 137-142.