INTERFACE: STRUMENTI INTERATTIVI PER L'ANIMAZIONE DELLE FACCE PARLANTI

Graziano Tisato, Piero Cosi, Carlo Drioli, Andrea Fusaro, Fabio Tesser

ISTC – SFD – CNR
Istituto di Scienze e Tecnologie della Cognizione
Laboratorio di Fonetica e Dialettologia

RIASSUNTO

Gli sviluppi recenti delle ricerche sui modelli di produzione e percezione della lingua parlata, così come sugli aspetti tecnologici dell'interazione uomo-macchina (riconoscimento della voce, sintesi di agenti conversazionali, insegnamento delle lingue, riabilitazione della voce, ecc.) richiedono l'acquisizione e l'elaborazione di grandi quantità di dati articolatori ed acustici [1, 2]. È noto infatti che questi dati si differenziano da lingua a lingua per la dimensione e la struttura dell'inventario fonologico. D'altra parte, la richiesta di questo tipo di dati è aumentata negli ultimi anni con il crescente interesse mostrato dalla comunità scientifica nel campo delle emozioni [3, 4].

Questo articolo presenta **InterFace**, un ambiente interattivo realizzato all'ISTC-SFD con lo scopo di facilitare tutte le fasi dell'elaborazione di questi dati [5].

Le caratteristiche di **InterFace** vorrebbero essere la semplificazione e l'automazione di gran parte delle operazioni, e l' integrazione in un ambiente di lavoro omogeneo delle applicazioni sviluppate all'ISTC-SFD per questo scopo.

InterFace permette di raggiungere tre principali finalità:

1. Estrarre dai dati acquisiti un certo insieme di misure dei parametri articolatori (apertura labiale, arrotondamento, protrusione, aggrottamento, asimmetrie labiali, ecc.) utili in campo fonetico e nello studio delle emozioni.
2. Ottenere da quegli stessi dati un modello parametrico ottimizzato dell'evoluzione dei parametri fonetici che tenga conto dei fenomeni di coarticolazione e che possa essere impiegato nelle Teste Parlanti (Talking Heads).
3. Sintetizzare il flusso dei dati audio-video necessari all'animazione di un agente conversazionale e capace di esprimere emozioni, nel nostro caso **Greta** [8], e **Lucia** [9].

Il sistema può maneggiare tre tipi di dati in ingresso:

1 - Dati reali acquisiti da ELITE (ELaboratore di Immagini Televisive) [7], che è un sistema optoelettronico per la cattura di andamenti cinematici di marker riflettenti la luce all'infrarosso.

Questi dati sono manipolati da quattro applicazioni scritte in Matlab:

- **Track**: permette la ricostruzione 3D delle traiettorie dei marker applicati sulla faccia di un soggetto. Consente inoltre di stabilire la corrispondenza fra punti acquisiti e punti definiti secondo un particolare standard di animazione (attualmente MPEG-4 [6]), e di ottenere in questo modo una tipica ***Data-Driven Synthesis*** [14].
- **Optimize**: utilizza i dati provenienti da Track per estrarre i coefficienti di articolazione fonetica, ottimizzati secondo un criterio di minimizzazione dell'errore da un modello di Cohen-Massaro [10], che è stato modificato per riprodurre i fenomeni di coarticolazione [9].
- **Apmanager**: consente di stabilire un certo numero di misure fra i dati articolatori di Track rispetto a punti, rette o piani di riferimento opportunamente definiti.
- **Mavis** (*Multiple Articulator VISualizer*, scritto da Mark Tiede agli ATR Research Laboratories, Tokyo [11]) che è usato per visualizzare e segmentare i parametri articolatori calcolati da APmanager.

2 – Dati simbolici XML ad alto livello, i quali sono elaborati da **AVengine** per produrre il flusso di dati audio-video di controllo dell'animazione facciale. Gli sviluppi più recenti a ISTC-SFD relativi all'*Affective Presentation Markup Language* (APML) [12] e al sistema di sintesi Festival [13] permettono ora di pilotare la sintesi audio con opportune etichette "emotive" a due differenti livelli: a livello più basso con i parametri definiti dalla *Voice Quality* (*spectral tilt, shimmer, jitter, aspiration noise*, ecc.), e ad un livello più elevato con una modellizzazione statistica dell'evoluzione dei parametri prosodici [15, 16]. Questo procedimento porta ad una ***Text-to-Animation Synthesis*** ovverosia ad una ***Symbolic-Driven Synthesis***.

3 – Dati a basso livello: sono creati dal programma **FacePlayer**, che consente l'*editing* di uno o più parametri di animazione MPEG-4 ed una immediata verifica del loro effetto con la sintesi video. Quest'ultima procedimento si può definire come una ***Manual-driven synthesis***.

Bibliografia

[1] Magno Caldognetto, E., Zmarich, C., Cosi, P., Ferrero, F. "Italian Consonantal Visemes: Relationship between Spatial/Temporal articulatory characteristics and coproduced acoustic signal", in *Proc. ESCA Workshop on Audio-Visual Speech Processing*, C. Benoît and R. Campell (Eds.), Rhodes, Greece, 1997, 5-8.

[2] Magno Caldognetto, E., Zmarich, C., Cosi, P., "Statistical definition of visual information for Italian vowels and consonants", *Internat. Conf. of Auditory-Visual Speech Processing - AVSP 1998*, Terrigal, Australia, 1998, (CD-Rom).

[3] Magno Caldognetto E., Cosi P., Drioli C., Tisato G., Cavicchio F., "Coproduction of Speech and Emotion: Bi-Modal Audio-Visual Changes of Consonant and Vowel Labial Targets", *Internat. Conf. of Auditory-Visual Speech Processing - AVSP 2003*, St. Jorioz, France, 2003, 209-214.

[4] Magno Caldognetto E., Cosi P., Drioli C., Tisato G., Cavicchio F., "Modifications of phonetic labial targets in emotive speech: Effects of the co-production of speech and emotions", special issue of *Speech Communication* (in press), 2004.

[5] Tisato, G., "InterFace: Interactive Tools for Facial Animation", http://www.csrf.pd.cnr.it/interface

[6] MPEG-4 standard. Home page: http://mpeg.telecomitalialab.com/standards/MPEG-4

[7] Ferrigno G., Pedotti A., "ELITE: A Digital Dedicated Hardware System for Movement Analysis via Real-Time TV Signal Processing", *IEEE Trans. on Biomedical Engineering*, BME-32, 1995, 943-950.

[8] Pelachaud C., Magno Caldognetto E., Zmarich C., Cosi P., "Modelling an Italian Talking Head", *Proc. AVSP 2001*, Aalborg, Denmark, 2001, 72-77.

[9] Cosi P., Fusaro A., Tisato G., "LUCIA a New Italian Talking-Head Based on a Modified Cohen-Massaro's Labial Coarticulation Model", *Proc. Eurospeech 2003*, Geneva, Switzerland, 2003, 127-132.

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

[11] Tiede, M.K., Vatikiotis-Bateson, E., Hoole, P. and Yehia, H, "Magnetometer data acquisition and analysis software for speech production research", ATR Technical Report TRH 1999, ATR Human Information Processing Labs, Japan, 1999.

[12] B. De Carolis, C. Pelachaud, I. Poggi, and M. Steedman. "Apml, a mark-up language for believable behavior generation", in H. Prendinger (Eds.), *Life-like Characters. Tools, Affective Functions and Applications*. Springer, 2004, 65-85.

[13] Taylor, P., Black, A., Caley, R., "The architecture of the Festival speech synthesis system", *3rd ESCA Workshop on Speech Synthesis*, 1998, 147-151. http://www.cstr.ed.ac.uk/projects/festival/

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

[15] Drioli C., Tisato G., Cosi P., Tesser F., "Voice Quality: Functions, Analysis and Synthesis", *Proc. of Voqual 2003, Voice Quality: Functions, Analysis and Synthesis, ISCA Workshop*, Geneva, Switzerland, 2003, 127-132.

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