IMPLEMENTAZIONE E VALUTAZIONE DI UNA FACCIA PARLANTE ESPRESSIVA IN UN SISTEMA DI E-TUTORING

EMANUELA MAGNO CALDOGNETTO

Istituto di Scienze e Tecnologie della Cognizione ISTC-CNR Padova
Via Martiri della Libertà 2, 35137, Padova
e-mail : emanuela.magno@pd.istc.cnr.it

FEDERICA CAVICCHIO

Center for Mind/Brain Studies, Università degli Studi di Trento
Corso Bettini, 31, 38060 Rovereto (Tn)
e-mail : federica.cavicchio@unitn.it

PIERO COSI

Istituto di Scienze e Tecnologie della Cognizione ISTC-CNR Padova
Via Martiri della Libertà 2, 35137, Padova
e-mail : piero.cosi@pd.istc.cnr.it

Riassunto

Il dibattito sul ruolo del tutor nell'e-learning è attualmente molto vivace e di grande importanza per chi, occupandosi di modellazione di interfacce uomo-macchina, vuole approcciare il problema dell'e-tutoring (Bevacqua *et al.,* in corso di stampa; Berry *et al.,* 2005; De Carolis*,* 2005). Partendo dalle conoscenze cognitive, linguistiche, fonetiche e tecnologiche necessarie per la messa a punto di un'interfaccia uomo-macchina e le conoscenze attuali sulla comunicazione multimodale umana, l'articolo illustra l'implementazione delle suddette informazioni e come queste, trasmettendo informazioni tramite segnali uditivi e visivi integrati, diano come risultato un'interfaccia più naturale e motivante nei processi di apprendimento nell'e-learning. Viene presentato il sistema di sintesi bimodale da testo che genera LUCIA, una Faccia espressiva parlante italiano messa a punto presso l'ISTC, Sezione di Padova. Sono infine illustrati i risultati di un test finalizzato alla valutazione della motivazione all'interazione con il sistema di tutoring da parte dello studente con i tre diversi tipi di interfaccia applicabili nell'e-learning: grafica, vocale e bimodale uditivo-visiva.

Abstract

The role of tutoring systems and the user modelling interface are central issues on in e-learning community studies (Bevacqua *et al.,* in press ; Berry *et al.,* 2005; De Carolis *et al.,* 2005). In this paper we propose a system based on a text-to-speech 3D synthesised animated face incorporating cognitive, linguistic and phonetics features (and the consequent multimodal integration of audio-visual signals) in an e-

learning context. The LUCIA (an expressive Talking in Italian Head) bimodal text-to-speech system by ISTC, Section of Padova, is described. Further we introduce the results of a test on motivation to interact with three different interfaces in HCI and e-learning: graphic interface, text-to-speech vocal synthesis interface and bimodal text-to-speech interface.

1. Il Ruolo del tutor nell'e-learning

In un sistema di e-learning il tutor assolve compiti diversi: deve creare una comunità didattica, guidare una discussione, risolvere possibili problemi all'interno di un gruppo, correggere eventuali errori di interpretazione dei materiali presentati, trasmettere valutazioni di esami (funzione di feedback), ecc. In questi compiti, espletati di solito in modalità asincrona, per definire la *social presence* (ovvero il senso di appartenenza a una comunità di discenti al fine di motivare la partecipazione), dovrà trasmettere anche stati affettivi ed emotivi, sulla base di scelte sociolinguistiche e paralinguistiche che determineranno il registro stilistico (p. es. formale/informale, autoritario/amichevole) con la conseguente scelta di tipologie frasali e lessicali o l'applicazione di *display rules* di espressione di stati d'animo ed emozioni verbali e non verbali.

1.1. Espressione delle emozioni nel corpus di interazioni didattiche

L'importanza dell'espressione delle emozioni (come in altri casi di CMC non a caso definiti *social interaction*) viene confermata dalle analisi del comportamento del docente nel momento dell'interazione faccia-a-faccia con lo studente (Poggi *et al.,* in corso di stampa). Se in un sistema e-learning il tutor utilizza solo messaggi scritti, per la trasmissione di emozioni, stati d'animo e atteggiamenti applicherà strategie già individuate dalle ricerche sugli SMS (Ursini, 2000; Tini Brunozzi e Danieli, in corso di stampa; Danieli, in corso di stampa) e confermate dalle nostre analisi su testi di chat e forum didattici (Magno Caldognetto *et al.,* 2005a; Magno Caldognetto *et al.,* in corso di stampa,a).

Nel corso del progetto FIRB MIUR "*Nuove tecnologie per la formazione permanente e reti nel sistema socioeconomico italiano*" abbiamo raccolto un corpus di interazioni sincrone e asincrone tra tutor e studenti (che si basano principalmente sulla CMC) allo scopo di indagare quali fossero le emozioni presenti e come fossero manifestate (Magno Caldognetto *et al.,* 2005a). In tali interazioni le classi di segnali linguistici e grafico-linguistici utilizzati per esprimere emozioni sono state: il lessico, le emoticon, la tipologia di saluti, la punteggiatura e l'intensificazione grafica, senza trascurare le loro possibili associazioni e co-occorrenze (esempi per i saluti: *bacioni/ bacionissimi/ un mare di baci; ciao, ciao!, CIAO, CIAO!!!, CIAOOOOOO!!!!!, !!!CIAOOOO!!!!!* ; *???? Daniela*). I principali segnali linguistici sono risultati nomi, aggettivi e verbi semanticamente connessi ad una emozione, ma anche indicatori morfologici (prevalentemente suffissi) e sintattici (frasi esclamative, dislocazioni a sinistra e fenomeni di focalizzazione), come già indicato per il parlato emotivo italiano da Poggi e Magno Caldognetto (2004).

Per quanto riguarda i segnali grafici, le emoticon sono state usate come sinonimi o rafforzativi del messaggio verbale che accompagnano[1]. L'emozione prevalentemente espressa da parte dei tutor e dagli studenti è risultata l'emozione sociale di *simpatia*, ma sono stati registrati casi di emozioni cognitive quali *rabbia e delusione* e, da parte di studenti, anche espressioni di emozioni di autoimmagine, *insoddisfazione e insicurezza* (emoticon :-/). I tre o più puntini di sospensione sono un altro segnale molto utilizzato nelle interazioni sincrone e asincrone e sembrano attribuire alla frase un significato contrapposto a quello che viene dato con il punto e il punto esclamativo, cioè il significato di asserire o comandare una cosa con un atteggiamento di dominanza, di assertività. I puntini sembrano dare una sfumatura semiotica esattamente opposta, mimando una intonazione "anassertiva", "non-conclusiva".

2. Implementazione di un tutor virtuale

Per superare le limitazioni della CMC (Baracco, 2002) di cui si è appena discusso e per assicurare maggiore accessibilità, usabilità e applicabilità ai sistemi e-learning, si ritiene che l'interfaccia grafica possa essere sostituita da interfacce vocali e bimodali. In particolare queste ultime dovrebbero garantire maggiore naturalezza, intelligibilità e appropriatezza e risultare maggiormente persuasive (Berry *et al.,* 2005; Morishima *et al.,* 2004) e motivanti all'interazione (Magno Caldognetto *et al.,* in corso di stampa).

Per creare tali Agenti, sulla base delle ricerche sulla generazione del linguaggio naturale e specificamente sul dialogo sviluppate soprattutto nell'ambito dell'IA e delle scienze cognitive, sono stati proposti diverse architetture e diversi formalismi che identificano e rappresentano le varie componenti del processo che, partendo dalla rappresentazione logica (e.g. la logica Bayesiana) e cognitiva delle conoscenze e dei goals / intenzioni del Mittente, genera messaggi nelle diverse modalità utilizzate dagli umani nell'interazione faccia-a-faccia (Cassell *et al.,* 1994, 2000; Castelfranchi, 2000; Poggi e Pelachaud, 2000; Pelachaud, 2003; De Rosis *et al.,* 2003; De Carolis *et al.,* 2000; De Carolis, 2005; Bevacqua e Pelachaud, in corso di stampa).

Questo processo comprende la specificazione del contenuto del dialogo, la pianificazione delle varie fasi del dialogo, la scelta (basata sulla conoscenza della comunicazione multimodale: cfr. Cassell, 2000; Magno Caldognetto e Poggi, 2001) dello strato informativo (lessicalizzazione di un'emozione o realizzazione vocale, per esempio) e del canale (acustico o visivo) su cui inviare le informazioni, anche sulla base delle ipotizzate conoscenze del Destinatatrio (Poggi e Magno Caldognetto, 1998; Poggi e Pelachaud, 2000; Cassell *et al*., 2000).

Problemi specifici sono stati proposti, tanto a livello teorico quanto tecnologico, dalla necessità che un Agente trasmetta correttamente emozioni adeguate agli scopi del dialogo (in connessione anche alla "sua" personalità) per assicurare il successo della diffusione di informazioni (Pelachaud *et al.,* 1996; Picard, 1998; Cassel *et al*., 2000; Poggi e Pelachaud, 2000; Poggi *et al*., 2000, 2004; Babu *et al.,* 2005; De Carolis, 2005).

[1] Nella CMC sincrona tra pari prevale invece l'uso autonomo: Magno Caldognetto *et al.,* 2004 e in corso di stampa.

Attualmente per una realizzazione soddisfacente, naturale di messaggi multimodali da parte di un Agente Virtuale si deve pianificare la coproduzione di parlato, visemi, visual prosody, sguardo, gesti, correlati acustici e visivi [2] delle emozioni.

Numerosi nell'ultimo decennio (per una revisione: Berry *et al.,* 2005) sono stati i prototipi di Agenti costituiti dalla sola Faccia Parlante: in questo caso le ricerche, pur implicando comunque una serie di conoscenze fondamentali linguistiche e fonetiche relative ad atti linguistici, strutture del dialogo, strutture semantiche, lessicali, sintattiche, morfologiche, fonologiche, quindi anche linguo-specifiche, si sono concentrate sui problemi tecnologici dei programmi per la sintesi del parlato da testo (TTS, Text To Speech), sulla coordinazione delle unità del parlato con i visemi e con la visual prosody, sui sistemi di animazione della Faccia (p.es. M-PEG4) e sulla riproduzione delle caratteristiche antropomorfe (e identitarie).

Più recentemente, per assicurare il successo di questi sistemi di interazione multimodale, nella formulazione dei programmi di dialogo si è dato spazio a regole comunicative sociali, p.es. regole di cortesia, scelta degli stili di parlato, trasmissione di stati affettivi ed emozioni, da cui dipenderà la valutazione della loro correttezza ed adeguatezza al contesto situazionale e quindi della loro naturalezza.

2.1. Ricerche dell'ISTC

2.1.1. Le ricerche analitiche e la raccolta dei dati

Per la creazione di una interfaccia uomo-macchina che funzioni come un tutor virtuale, soddisfacendo le esigenze sociali richieste dall'e-learning, è stata utilizzata la Faccia Parlante LUCIA, un sistema TTS adattato all'italiano (Cosi *et al..* 2001, 2002c).

Al fine di assicurare l'intelligibilità e la naturalezza della Faccia Parlante e garantire la robustezza della trasmissione multimodale, nell'implementazione di LUCIA sono stati considerati e utilizzati i risultati di una serie di ricerche svolte presso la Sezione di Padova dell'ISTC sull'italiano parlato e sul parlato emotivo riguardanti:

- **le caratteristiche cognitive e linguistiche del parlato emotivo** (Poggi e Magno Caldognetto, 2004): sono state elencate ed esemplificate le diverse risorse linguistiche per esprimere emozioni, cioè le risorse lessicali, sintattiche (tipologie frasali, focalizzazione), morfologiche e fonologiche, e le loro possibili combinazioni.
- **il lessico delle emozioni:** sono state indagate (D'Urso *et al.,* in corso di stampa), in termini di distanza psicologica, le relazioni tra termini emotivi che gravitano intorno alle emozioni *gioia* e *tristezza,* testando le relazioni tra una serie di aggettivi, p.es. *felice, contento, lieto, divertito, gioioso, lieto* e *triste, disperato, malinconico, addolorato, rammaricato,* ecc.. Con lo stesso metodo è stata inoltre studiata la gradazione di intensità delle due emozioni di partenza ottenuta con modificatori quantitativi quali suffissi o avverbi (p.es. "molto felice", "abbastanza felice", "felicissimo", ecc.). I risultati di questo studio hanno permesso di stabilire delle corrispondenze tra diversi mezzi linguistici per

[2] Anche le informazioni sulla personalità e l'atteggiamento (ovvero la conservazione di una emozione su lungo periodo) sono oggetti di modellazione dell'interfaccia (Ball e Breese, 2000).

definire rapporti all'interno di una stessa famiglia di emozioni e potranno essere utilizzati nel sistema di sintesi bimodale per definire le modifiche delle caratteristiche acustiche e visive di diversi stati emotivi, ottenute intensificando o diminuendo quelle individuate per le emozioni di base.

- **I correlati acustici delle emozioni:** sono state analizzate le modificazioni macro- e micro-prosodiche indotte dall'espressione delle emozioni rispetto al parlato non emotivo (Magno Caldognetto, 2002; Magno Caldognetto e Ferrero, 1996; Magno Caldognetto *et al.,* 1998a), anche in un'ottica cross-culturale (Kori e Magno Caldognetto 1992).

- **i visemi dell'italiano:** sulla base di test percettivi visivi sono stati individuati i visemi, cioè le classi di movimenti articolatori visibili che trasmettono la stessa informazione fonologica, dell'italiano, sia vocalici che consonantici[3]. Per questi ultimi è stata anche studiata la loro variabilità a seconda del contesto vocalico e della staticità e dinamicità dello stimolo (Magno Caldognetto e Vagges, 1990; Magno Caldognetto e Zmarich, 2001). Analisi eseguite con sistemi optoelettronici (ELITE) hanno portato alla quantificazione articolatoria dei bersagli vocalici e consonantici in termini di parametri fonetico-fonologici quali Apertura Labiale, Larghezza Labiale, Protrusione labbro Superiore e Inferiore, alla definizione delle modificazioni dovute alla coarticolazione (Magno Caldognetto e Zmarich, 1998, 1999) e all'individuazione della Larghezza Labiale quale indice più rilevante per il riconoscimento percettivo visivo (Magno Caldognetto *et al.,* 1997, 1998b).

- **Le relazioni tra unità segmentali acustiche del parlato e movimenti labiali**: sono state indagate la coerenza, la sincronia o l'anisocronia tra i due segnali (Magno Caldognetto *et al.*, 1997), importanti per la teoria acustica di produzione del parlato, per le teorie fonologiche articolatorie (Browman e Goldstein, 1995), per le regole del sinergismo percettivo e le teorie sulla percezione dei suoni linguistici (Erber, 1975; Summerfield, 1987; Massaro, 1987, 1996; Fowler e Rosenblum, 1991; Best, 1995; ma anche McGurk e Mac Donald, 1976).

- **Le regole di interazione tra i movimenti labiali che realizzano unità segmentali linguistiche e le configurazioni labiali che caratterizzano le emozioni nel parlato emotivo** (Figura 1).

[3] La trasmissione di informazione fonologica tramite movimenti articolatori visibili, particolarmente importante in condizioni ambientali di rumore e in caso di patologie dell'udito, è tradizionalmente conosciuta come Lettura Labiale, Lip Reading o Speech Reading (Dodd e Campbell, 1987; Campbell *et al.,* 1998).

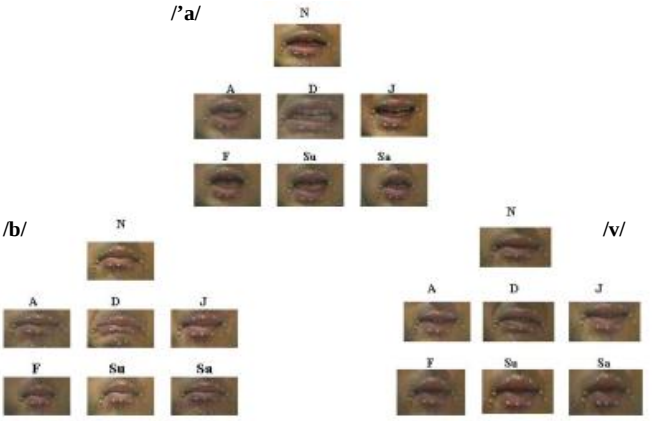

Figura 1 *— Modificazioni della configurazione labiale dovuto alle sei emozioni di base (A=anger, D=disgust, J=joy, F=Fear, SU=surprise, Sa=Sadness) in contesto articolatorio /'a/, /b/, /v/.*

Le ricerche eseguite con ELITE (Magno Caldognetto *et al*., 2004a, 2004b, 2005b; cfr. anche Nordstrand et *al*., 2004) hanno evidenziato come le configurazioni labiali emotive caratteristiche di gioia, paura, collera, sorpresa, tristezza, disgusto, modifichino significativamente i valori dei parametri (per l'elenco dei parametri presi in considerazione vedi Figura 2) rispetto alla produzione non emotiva, ma non si sovrappongano semplicemente ai bersagli articolatori labiali vocalici e consonantici della produzione non emotiva, piuttosto interagiscano con i parametri fonetici che specificano le unità linguistiche in modo da preservare il loro ruolo fonologico a seconda della tipologia della consonante o vocale. Se i valori di un parametro sono distintivi, come nel caso dell'Apertura Labiale, le modificazioni imposte dalla realizzazione delle configurazioni delle varie emozioni rispettano la diversità fonologica dei gradi di costrizione che distinguono, per esempio, le consonanti occlusive bilabiali dalle costrittive labiodentali, mentre nella realizzazione della vocale aperta centrale /a/ i valori dell'Apertura Labiale presentano un range di variazione molto ampio e distinto soprattutto per emozione (Magno Caldognetto *et al.,* 2004a, 2004b). Per quanto riguarda il parametro di Larghezza Labiale, all'interno dei valori di maggiore appiattimento registrati per le emozioni di gioia e disgusto, viene evidenziata la diversità significativa tra i valori rilevati per la consonante /v/, molto vicini a quelli della produzione neutra caratterizzata articolatoriamente da una rima a fessura, rispetto a quelli di /b/ e /a/.

La quantificazione degli spostamenti verticali degli angoli delle labbra e la loro eventuale asimmetria ha invece dimostrato che questi parametri caratterizzano esclusivamente le emozioni, in particolare gioia e disgusto.

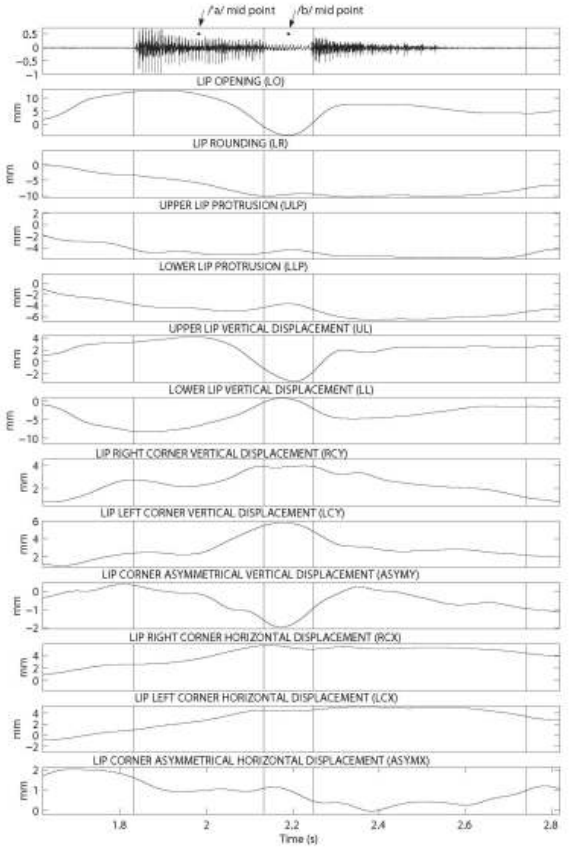

Figura 2*— Parametri di analisi articolatoria del parlato emotivo. L'immagine si riferisce in particolare agli andamenti articolatori dei parametri durante la produzione della parola /'aba/ pronunciata con emozione di disgusto.*

-**L'individuazione dei movimenti e delle configurazioni facciali correlati alla *visual prosody***: innalzamento delle sopracciglia, apertura degli occhi, corrugamento della fronte, movimenti della testa, ecc. e il loro rapporto semantico e temporale con le unità soprasegmentali del parlato dipendenti principalmente dalla struttura topic-comment frasale (cfr. Beskow, 1997; Beskow *et al.,* 2004; Beskow e Cerrato, in corso di stampa) sono stati studiati grazie ad un programma di segmentazione ed etichettatura di segnali multimodali, ANVIL, in cui è stata implementata la Partitura, un sistema per l'analisi e la trascrizione di parlato, prosodia, movimenti facciali (della bocca, degli occhi, delle sopracciglia), della testa e del corpo, e gesti delle mani, che prevede per ciascuna modalità comunicativa diversi livelli di analisi: descrizione, tipologia, significato, funzione semantica (Poggi e Magno Caldognetto, 1996; Magno Caldognetto *et al.,* 2004c).

E' evidente che la naturalezza della Faccia Parlante e la sua adeguatezza a compiti comunicativi che prevedano *social presence* dipenderà dalla correttezza con cui gli indici acustici e visivi che veicolano tutte queste informazioni linguistiche e

paralinguistiche (Magno Caldognetto *et al*., in corso di stampa, b) risulteranno coprodotti nel flusso dei segnali generati dal sistema di sintesi bimodale.

2.1.2. La sintesi bimodale da testo (*TXT2animation)*

A partire da questi dati è stato costruito il materiale per gli esperimenti di sintesi audio-visiva da testo eseguiti con la Faccia Parlante LUCIA (Cosi *et al.,* 2002a; Cosi *et al.,* in corso di stampa). La sintesi audiovisiva da testo (*TXT2animation*) partendo da un testo etichettato emotivamente (p. es. con APML: De Carolis *et al*., 2003; De Rosis *et al.,* 2003), lo trasforma nel corrispondente emotivo per la Faccia Parlante associando un file audio di tipo WAV, sintetizzato tramite la versione italiana di FESTIVAL (Cosi *et al.* 2001), modificato per ottenere le adeguate espressioni emotive (Drioli *et al.,* 2003; Cosi *et al.,* 2002b, 2002c; Tesser *et al.,* 2005), a un set di FAP (*Facial Action Points),* ottimizzati da un modello di coarticolazione labiale e da specifici movimenti facciali ricavati per ogni emozione tramite regole ricavate da addestramento a partire da un database di espressioni facciali emotive (cfr. sistema INTERFACE: Tisato *et al*., 2004, 2005; Figura 3). Ogni particolare sequenza di azioni facciali emotive viene generata deformando il modello facciale a partire dallo stato in cui la faccia si presenta neutra, priva di emozioni (Cosi *et al.*, 2004).

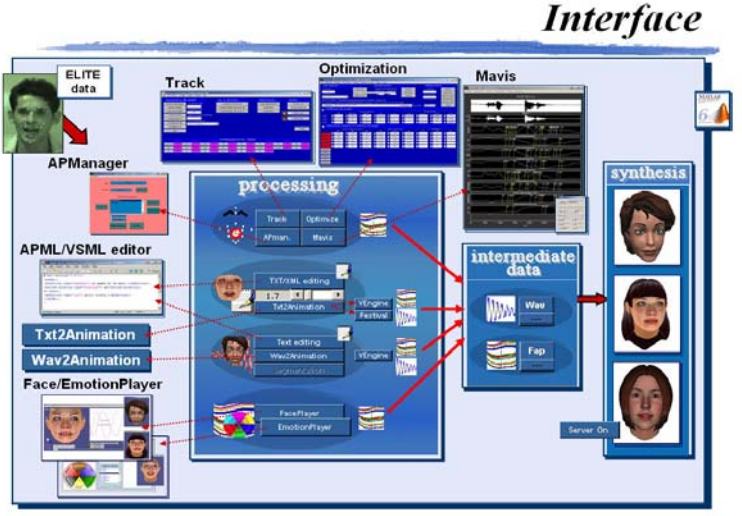

Figura 3 *— Il programma InterFace con le tre aree funzionali: elaborazione, editing dei dati audio-visuali e animazione con Lucia e Greta.*

3. Test confrontativo tra le interfacce

Questo test è stato eseguito allo scopo di fare una prima indagine sulla *motivazione* all'interazione con i sistemi informatici comunemente usati nelle piattaforme e-learning e in particolare nei sistemi informatici per l'interazione tra studenti e tutor on line.

La diffusione sempre più ampia dei *tutoring systems* porta a compiere una riflessione sulle tipologie di interfacce uomo-macchina usate nei sistemi di e-learning e a verificarne l'accettabilità e l'usabilità da parte degli utenti. Con il progresso delle applicazioni tecnologiche, alle interfacce grafiche comunemente utilizzate sono state affiancate per l'e-learning interfacce vocali basate su sintesi da testo (pensate soprattutto per i disabili) e interfacce bimodali uomo macchina (Avatar o Facce Parlanti).

Negli ultimi anni nel campo dello studio della comunicazione mediata da computer (CMC) sono stati compiuti numerosi studi che hanno smentito progressivamente il presupposto che l'interazione tra utenti online fosse un tipo di comunicazione "fredda". In particolare studi sulle interazione tra utenti in chat e forum hanno dimostrato che si tratta di un tipo di comunicazione particolarmente ricca di espressività, basti pensare nelle interfacce scritte al grande uso degli emoticon sia per rafforzare che per disambiguare il messaggio (Spears *et al*., 2001). Proprio la difficoltà di interpretare univocamente i segnali fa presupporre che il passaggio da un tipo di comunicazione "unimodale", basata sulla videoscrittura, a una "multimodale" come quella della Faccia Parlante sia in grado di aumentare il gradimento delle interazioni in rete e soprattutto motivi gli utenti ad interagire a distanza facilitarndo così la *social presence*, fattore di primaria importanza per la buona riuscita dell'apprendimento on line (Garrison *et al.,* 2000).

Per comprendere se questa ipotesi sia valida e se la motivazione all'interazione venga o meno facilitata in un task di e-learning dalla tipologia di interfaccia è stato allestito un test che confronta tre diverse tipologie di interfaccia: grafica, sintesi vocale e sintesi bimodale (Faccia Parlante LUCIA, ISTC CNR di Padova). Oltre alla motivazione questo esperimento si propone di misurare l'impatto della componente emozione (positiva o negativa) sulla volontà espressa dai soggetti di continuare l'interazione con una delle tipologie di interfaccia. L'ipotesi alla base della ricerca sostiene che i partecipanti si sentono, a parità di emozione trasmessa, più motivati all'interazione con il sistema di sintesi bimodale, dal momento che si trovavano di fronte ad un volto, seppur artificiale, che comunicava con loro trasmettendo emozioni. Nel caso però che l'interfaccia bimodale risponda loro con espressioni ed emozioni negative, si è previsto che la motivazione ad interagire sarebbe stata minore rispetto alle altre tipologie di interfacce, a parità di emozione trasmessa. Nessuna predizione era invece stata formulata per l'espressione non emotiva, neutra.

Per confrontare le tre tipologie di interfacce al meglio delle loro possibilità espressive e quindi per limitare l'impatto negativo dipendente dal possibile diverso livello tecnologico (Berry *et al.,* 2005), il messaggio grafico è stato presentato nella forma più completa, accompagnato cioè da segnali ortografici ed emoticon, mentre tanto la sintesi vocale quanto quella bimodale non sono state prodotte sulla base di una sintesi da testo, ma di un sistema di risintesi (Data driven Synthesis in Tisato *et al.* 2004, 2005). L'animazione della Faccia Parlante LUCIA è stata infatti ottenuta tramite un sistema che passa come input al sistema MPEG 4 i *files* provenienti dai rilevamenti dei movimenti labiali, mandibolari e facciali ottenuti dal sistema optoelettronico ELITE, effettuati da un attore (Figura 4).

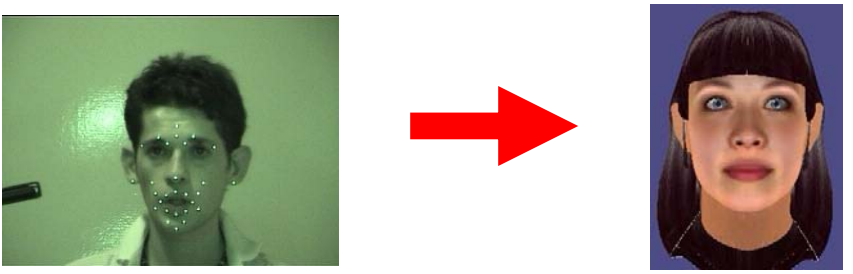

Figura 4 *— Sistema di risintesi della Faccia Parlante LUCIA.*

Allo stesso modo inoltre è stato utilizzato sia per l'interfaccia bimodale che per quella vocale un sistema di risintesi del parlato ottenuto partendo da un file audio invece che da un testo[4]. Per un esempio rimandiamo al filmato seguente: http://www.pd.istc.cnr.it/LUCIA/Movies/FIRB-test/Lucia-R3P04-Lucia-3P01-SalveSorridente.avi

3.1. Metodo

Ai partecipanti (42 soggetti, di cui 12 maschi e 30 femmine, studenti universitari provenienti da facoltà diverse, di età compresa tra i 19 e i 30 anni) venivano presentate in sequenza randomizzata due modalità di interazione tenendo costante il tipo di emozione presentata dall'interfaccia. Infatti nell'ipotesi iniziale si voleva verificare se la tipologia di interfaccia e l'emozione trasmessa andassero a modificare la motivazione all'interazione con la data interfaccia utente. L'interfaccia, dopo una breve introduzione, poneva all'interagente una domanda di cultura generale. Il partecipante dava una risposta di tipo si/no dopo la quale l'interfaccia dava una valutazione al partecipante.

Lo schema dell'interazione era, dunque, il seguente[5]:

Tutoring system saluto (neutro o *affective*, cioè con emozione, *"salve"*)

Tutoring system somministrazione di informazione (es. *"Le vette più alte d'Italia sono il Monte Bianco, il Monte Rosa e il Monte Cervino"*) http://www.pd.istc.cnr.it/LUCIA/Movies/FIRB-test/Lucia-R3P04-Lucia-R2P04-LeVetteEnunciativa.avi

Tutoring system domanda (es. "*Il Monte Cervino è alto 3827 metri*?")[6] http://www.pd.istc.cnr.it/LUCIA/Movies/FIRB-test/Lucia-R3P04-Lucia-R2P07-MonteCervinoEnunciativa.avi

Studente risposta sì/no cliccando su tasto a schermo.

---

[4] Un algoritmo di segmentazione automatica basato su un sistema ASR per l'italiano estrae i confine dei fonemi dai quali è generate l'animazione facciale.

[5] Questo schema di interazione è stato mantenuto costante per tutti e tre i tipi di interazione, quindi sia per l'interfaccia scritta che per quella vocale o bimodale uditivo-visiva.

[6] Le domande poste ai partecipanti erano volutamente ambigue per non generare contestazione rispetto alle valutazioni ricevute dal sistema informatico.

Tutoring system valutazione positiva o negativa (produzione neutra o *affective* positiva o negativa, *"bene la risposta è corretta"* http://www.pd.istc.cnr.it/LUCIA/Movies/FIRB-test/Lucia-R2P05-BeneSorridente.avi con emozione positiva oppure "peccato la risposta è sbagliata" http://www.pd.istc.cnr.it/LUCIA/Movies/FIRB-test/Lucia-R2P09-PeccatoSorridente.avi con emozione positiva empatica o negativa di rabbia trattenuta[7] http://www.pd.istc.cnr.it/LUCIA/Movies/FIRB-test/Lucia-R2P02-PeccatoArrabbiata.avi

Al soggetto veniva presentato questo schema di interazione con due diverse tipologie di interfaccia, cambiando la domanda ma mantenendo inalterata l'emozione nella valutazione espressa dall'interfaccia (neutra, *affective* positiva o negativa). Di conseguenza le possibilità di risposta del soggetto erano solo alternative (esatta/sbagliata, si/no), mentre le domande e le valutazioni del tutor erano alternative e previste.

Dopo l'interazione con il sistema informatico il partecipante rispondeva ad un questionario. L'esperimento è stato condotto in una stanza in cui vi era un computer, posto sopra un tavolo, con cui lavorava il partecipante.

3.2. Risultati

Per valutare la preferenza/impatto motivazionale dei 3 tipi di interfaccia si riportano i risultati delle medie dei punteggi ottenuti delle tre espressioni emotive proposte. I punteggi attribuiti erano selezionati in una scala da 1 a 5 ed erano relativi alla volontà espressa dai partecipanti a continuare l'interazione (variabili "continuare" e "interagire") e alla comprensione del messaggio veicolato dall'interfaccia. Considerando le statistiche ANOVA ottenute incrociando tra loro le tipologia di interfaccia (grafica, sintesi vocale e bimodale, vedi Tabella 1) si presenta una differenza significativa solo quando le modalità di sintesi vocale e grafica sono in relazione con la variabile dipendente "***soddisfazione***" all'interazione (p=0.003). Quando le sintesi vocale e bimodale sono presentate insieme non c'è significatività, il che fa pensare ad una assenza di differenza nell'ordine di presentazione delle due interfacce.

Tabella 1 — *Risultati della Anova a due fattori e significatività delle risposte "soddisfazione" e "comprensione" in relazione alle modalità di interfaccia.*

	Vocale-grafico *Grafico-vocale*	*Bimodale-vocale* *Vocale-bimodale*	*Grafico-bimodale* *Bimodale-grafico*
SODDISFAZIONE	0.003	-	-
COMPRENSIONE	0.035	0.035	0.040

In relazione alla variabile dipendente "***comprensione***" notiamo una differenza significativa (p=0.035) in presenza dell'interfaccia con sintesi vocale rispetto a quella

[7] Per una collocazione semantica dei termini emotivi si rimanda a D'Urso et al. In corso di stampa.

con sintesi bimodale, come pure tra interfaccia scritta e bimodale, facendo risaltare l'interfaccia vocale come la meno comprensibile delle tre. Infine c'è una differenza significativa (p=0.040) quando le interfacce considerate sono bimodale e grafica.

Per quanto riguarda la componente emozione, va sottolineato che per l'interfaccia grafica sono state implementate due risposte emotive oltre a quella neutra, la risposta definita "*bene empatico*" e quella definita "*peccato empatico*", mentre non è stata individuata un'emoticon che esprimesse in modo soddisfacente la rabbia trattenuta.

Per quanto riguarda quindi queste due espressioni emotive i risultati per l'interfaccia *grafica* sono i seguenti[8]:

Bene Empatico: la modalità grafica ha ottenuto punteggi medi piuttosto bassi (da 1.5 a 2.5), tranne in un caso in cui il testo scritto è stato presentato come prima modalità di interazione, ottenendo un punteggio di 5. I punteggi sono leggermente (0.50 in media) più elevati quando la modalità grafica viene presentata in prima posizione. La comprensione del messaggio è molto alta (9.5-10); questo probabilmente è dovuto al livello scolare elevato dei partecipanti che erano molto abituati alla forma scritta della presentazione a schermo.

Peccato Empatico: sono stati ottenuti punteggi tra 2.5 e 4. Non c'è una sostanziale differenza a seconda della posizione in cui l'interfaccia a caratteri viene presentata.

Per l'interfaccia *vocale*, come per quella bimodale, si aggiungono i risultati dell'emozione di "*rabbia trattenuta*", cioè non tanto una rabbia centrale, ma spostata verso l'irritazione. I tre tipi di risposta emotiva hanno ottenuto i seguenti risultati:

Bene Empatico: i punteggi sono compresi tra 2.5 e 4. Non si rilevano grandi differenze per quanto riguarda la posizione di presentazione.

Peccato Empatico: ha ottenuto punteggi compresi tra 4 e 3 (tranne in un caso, dove abbiamo 0.5). Si registrano dei valori leggermente più bassi (1.5 di media) quando la modalità vocale è presentata come seconda tipologia di interfaccia.

Peccato Arrabbiato: il punteggio è tra 4 e 2 ed è più basso quando l'interfaccia vocale viene presentata con questa emozione come seconda modalità di interazione.

L'interfaccia *bimodale* ha ottenuto i seguenti punteggi:

Bene Empatico: la modalità bimodale ha dei punteggi per "continuare" e "interagire" tra il 3 e il 4 (su una scala di 5), soprattutto quando è presentata in maniera incrementale (cioè come seconda modalità).

Peccato Empatico: l'interfaccia bimodale ottiene punteggi più bassi (tra 0.5 e 3) rispetto alla risposta positiva, sia che venga presentata come prima modalità o come seconda modalità e viene ritenuta più efficace la modalità scritta.

---

[8] Nella scala attributiva per la domanda "*continuare"* il punteggio 1 equivale a "Sì, ancora 1/2 volte", 2 a "Sì, ancora 3/4 volte, 3 a "Sì, anche 5 volte o più", 4 a "No, basta"; per la domanda numero 4, "*interagire*", il punteggio 1 equivale a "Per niente", il punteggio 2 a "Un po'", 3 ad "Abbastanza", 4 a "Molto", 5 a "Moltissimo".

Peccato Arrabbiato: l'interfaccia bimodale presenta punteggi alti in entrambi i casi (punteggi compresi tra 3 e 4); i punteggi sono più alti quando viene presentata come seconda modalità di interazione.

4. Conclusioni

I risultati del test hanno fatto sorgere dei dubbi riguardo la nostra ipotesi iniziale, quella secondo cui un'interazione con la sintesi bimodale sarebbe stata preferita dai partecipanti all'esperimento. Si pensava che l'interazione con un volto sarebbe risultato maggiormente interessante e stimolante per chi partecipava alla ricerca. Nella nostra ipotesi, però, non era stato considerato il fattore "famigliarità": è necessario cioè tenere conto che nella quotidianità riceviamo e trasmettiamo messaggi in continuazione, ma le "modalità" di cui usufruiamo sono acustiche e/o scritte. In particolare di fronte a un computer si ha a che fare con interfacce scritte. Sono rari, se non assenti, i casi in cui ci troviamo ad interagire con una sintesi bimodale simile a quella presentata nell'esperimento, mentre assai più numerose sono le possibilità di interagire con una sintesi vocale, per esempio attraverso un altro sistema di comunicazione, ossia il telefono.

I risultati indicano che l'interfaccia bimodale è più apprezzata quando viene presentata come seconda modalità di interazione, e questo accade perché il partecipante si è ambientato al contesto, ha avuto modo di inserirsi nella situazione e, quindi, è più "pronto" ad interagire con un'interfaccia bimodale. Sempre per quanto riguarda l'interfaccia bimodale, notiamo dai dati raccolti una variazione significativa per *emozione*, piuttosto che per posizione: i punteggi più bassi risultano nelle risposte "peccato arrabbiato" e "peccato empatico" rispetto alla risposte "bene empatico". Questa osservazione ci fa pensare che, come nella nostra ipotesi iniziale, sia poco gratificante rendersi conto di aver sbagliato la risposta, soprattutto se questa consapevolezza ci deriva da un volto artificiale che interagisce con noi coinvolgendoci e trasmettendoci emozioni (Poggi, in corso di stampa). Al contrario, risulta piacevole sentire, e vedere, confermata la correttezza della risposta marcata emotivamente. Infine non si sono notate grandi differenze tra "peccato arrabbiato" e "peccato empatico", probabilmente perché l'interfaccia vocale non ha proposto l'emozione di rabbia centrale, ma le sue caratteristiche erano spostate verso l'irritazione, la rabbia trattenuta, che quindi è risultata poco riconoscibile presentata solo attraverso il canale acustico.

Concludendo, i risultati di questo primo esperimento mostrano che un setting sperimentale interattivo e una sintesi bimodale molto buona (ottenuta, ricordiamo, con risintesi) ottengono punteggi buoni, ma meno buoni rispetto a una forma di espressione consolidata di comunicazione come la scrittura e, in particolar modo, la videoscrittura.

Ancora aperta e in corso di discussione, anche alla luce dei risultati qui portati, è il ruolo delle emozioni nell'interazione con interfacce bimodali che risulta essere cruciale per l'accettazione di una interfaccia come naturale e motivante.

Bibliografia

Anvil Home Page: **http://www.dfki.de/~kipp/anvil/ .**

Ball G., Breese J. (2000). Emotion and Personality. In Cassell J., Sullivan J., Prevost S., Churchill E. (Eds.), *Embodied Conversational Agents*. Cambridge (Mass.): MIT Press. 189-219.

Babu S., Schmugge S., Inugala R., Rao S., Barnes T., Hodges L.F. (2005). Marve: A Prototype Virtual Human Interface Framework for Studying Human-Virtual Human Interaction. In Panayiotopoulos T., Gratch J., Aylett R., Ballin D., Olivier P., Rist T. (Eds.), *Intelligent Virtual Agents*, Proceedings of *5th International Working Conference IVA 2005*.121-133.

Baracco A. (2002). La comunicazione mediata dal computer, in C. Bazzanella C. (Ed.), *Sul dialogo. Contesti e forme di interazione verbale*, Milano: Edizioni Angelo Guerini e Ass. 253-267.

Berry D. C., Butler L.T., de Rosis F. (2005). Evaluating a Realistic Agent in an Advice-Giving Task, *International Journal of Human-Computer Studies*, 63, 304-327.

Beskow J. (1997). Animation of Talking Agents, in Proceedings of *ESCA Workshop on Audio-visual Speech Processing*. Rhodes (Greece). 149-152.

Beskow J., Cerrato L., Granstrom B., House D., Nordstrand M., Svanfeldt G. (2004). Expressive Animated Agents for Affective Dialogue Systems. In André E., Dybkjaer L., Minker W., Heisterkamp P. (Eds.), Proceedings of *ADS 2004- Tutorial and Research Workshop on Affective Dialogue Systems*. Berlin: Springer Verlag. 240-244.

Beskow J., Cerrato L. (in corso di stampa). Evaluation of the Expressivity of a Swedish Talking Head in the Context of Human–Machine Interaction. In Magno Caldognetto E., Cavicchio F. (Eds.), *Atti del 1° Convegno Nazionale GSCP (Gruppo di Studio della Comunicazione Parlata) su Comunicazione parlata e manifestazione delle emozioni (Padova 30/10-1 /11 2004).* Napoli: Liguori Editore.

Best C. T. (1995). A Direct Realist View of Cross-Language Speech Perception. In Strange W. (Ed.), *Speech Perception and Linguistic Experience. Issues in Cross-Languages Research*. Baltimore USA: York Press. 171-204.

Bevacqua E., Pelachaud C. (in corso di stampa). ECAs Espressivi. In Magno Caldognetto E., Cavicchio F. (Eds.), Atti del 1° Convegno Nazionale GSCP (Gruppo di Studio della Comunicazione Parlata) su *Comunicazione Parlata e manifestazione delle emozioni"* **(**Padova 30/10-1/11 2004). Napoli: Liguori Editore.

Bevacqua E., Mancini M., Peters C., Pelachaud C., Ochs M., Ech Chafai N. (in corso di stampa). Abilità socio-emotive per Agenti Virtuali dedicati all'e-learning. In Magno Caldognetto E., Cavicchio F. (Eds.), *Aspetti emotivi e relazionali nell'e-learning*. Firenze: FUP Florence University Press.

Browman C.P., Goldstein L. (1995). Dynamics and Articulatory Phonology. In Port R. F., Van Gelder (Eds.), *Mind as Motion*. Cambridge (Mass): MIT Press. 175-193.

Campbell R., Dodd B., Burnham D. (Eds.) (1998). *Hearing by Eye II*. Hove UK: Psychology Press**.**

Castelfranchi C.(2000). Affective Appraisal vs Cognitive Evaluation in Social Emotions and Interactions. In Paiva A. (Ed.), Affective Interactions. Towards a New Generation of Computer Interfaces. Berlin: Springer-Verlag.76-106.

Cassell J. (2000). Nudge Nudge Wink Wink: Elements of Face-to-Face Conversation for Embodied Conversational Agents, in Cassell J., Sullivan J., Prevost S., Churchill E. (Eds.), *Embodied Conversational Agents*. Cambridge (Mass.): MIT Press. 1-27.

Cassell J., Bickmore T., Campbell L., Vilhjalmasson H., Yan H. (2000). Human Conversation as a System Framework: Designing Embodied Conversational Agents, in J. Cassell, J.Sullivan, S.Prevost, E. Churchill (Eds.), *Embodied Conversational Agents*. Cambridge (Mass.): MIT Press. 29-63.

Cosi P., Tesser F., Gretter R., Avesani C. (2001). Festival Speaks Italian!. In Proceedings of *EUROSPEECH 2001*. Aalborg, Denmark. 509-512.

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

Cosi P., Tesser F., Gretter R., Pianesi F. (2002 b). A Modified 'PaIntE Model' for Italian TTS. In Proceedings of *IEEE Workshop on Speech Synthesis*. Santa Monica, California. CDROM.

Cosi P., Avesani C., Tesser F., Gretter R., Pianesi F. (2002 c). On the Use of Cart-Tree for Prosodic Predictions in the Italian Festival TTS. In Cosi P., Magno Caldognetto E., Zamboni A. (Eds.), *Voce, Canto, Parlato. Studi in onore di Franco Ferrero*. Padova: UNIPRESS. 73–81.

Cosi P., Fusaro A., Grigoletto D., Tisato G. (2004). Data Driven Tools for Designing Talking Heads Exploiting Emotional Attitudes. In André E., Dybkjaer L., Minker W. , Heisterkamp P. (Eds.), *Affective Dialogue Systems, Tutorial and Research Workshop, ADS 2004*. Berlin: Springer–Verlag. 101-112.

Cosi P., Drioli C., Fusaro A., Tesser F., Tisato G. (in corso di stampa). EMOTIONPLAYER: dalla teoria alla pratica. In Magno Caldognetto E., Cavicchio F., Cosi P. (Eds.), Atti del 1° Convegno Nazionale GSCP (Gruppo di Studio della Comunicazione Parlata) su *Comunicazione Parlata e Manifestazione delle Emozioni* **(**Padova 30/10-1/112004). Napoli: Liguori Editore.

Danieli M. (in corso di stampa). Segmenti di discorso e profili intonativi. In Magno Caldognetto E., Cavicchio F. (Eds.), *Aspetti emotivi e relazionali nell'e-learning*. Firenze: FUP Florence University Press.

De Carolis B. (2005). My Tutor : A Personal Tutoring Agent. In Panayiotopoulos T., Gratch J., Aylett R., Ballin D., Olivier P., Rist T. (Eds.), *Intelligent Virtual Agents*, Proceedings of 5th International Working Conference IVA 2005, 478-488.

De Carolis B., Pelachaud C., Poggi I., Steedman M. (2003). APML, a Markup Language for Believable Behavior Generation. In Prendiger H., Ishizuka M. (Eds.), *Life–like Characters. Tools, Affective Functions and Applications*. Berlin: Springer–Verlag. 65-86.

de Rosis F., Pelachaud C., Poggi I., Carofiglio V., De Carolis B. (2003). From GRETA's Mind to her Face: Modelling the Dynamics of Affective States in a Conversational Embodied Agent. *International Journal of Human-Computer Studies. Special Issue on Applications of Affective Computing in HCI.* 59, 81-118.

Dodd B., Campbell R. (Eds.) (1987). *Hearing by Eye: The Psychology of Lip-Reading*. London : LEA.

Drioli C., Tisato G., Cosi P., Tesser F. (2003). Emotions and Voice Quality: Experiments with Sinusoidal Modelling. Proceedings of *VOQUAL ESCA Workshop*. Geneva, Switzerland. 127-132.

D'Urso V., Cavicchio F. Magno Caldognetto E., (in corso di stampa). Le etichette lessicali nelle ricerche sperimentali sulle emozioni: problemi teorici e metodologici. In Magno Caldognetto E., Cavicchio F. (Eds.), Atti del 1° Convegno Nazionale GSCP (Gruppo di Studio della Comunicazione Parlata) su *Comunicazione Parlata e Manifestazione delle Emozioni (*Padova 30/10-1/11 2004). Napoli: Liguori Editore.

Erber N.P. (1975). Auditory-Visual Perception of Speech. *Journal of Speech and Hearing Disorders*. 40, 481-492.

Fowler C. A., Rosenblum L.D. (1991). The Perception of Phonetic Gestures. In Mattingly I.G., Studdert-Kennedy M. (Eds.), *Modularity and the Motor Theory of Speech Perception*. Hillsdale, N.J.: Lawrence Erlbaum Ass.. 33-59.

Garrison D.R., Anderson T., Archer W. (2000) Critical enquiry in a text-based environment: Computer conferencing in higher education. *The Internet and higher education,* 2(2-3), 1-19.

Kori S., Magno Caldognetto E. (1991). Cross-Cultural Perception of Emotions through Synthetic Vowels. Proceedings of the *12th International Congress of Phonetic Sciences*. Aix-en-Provence. Vol.3, 310-313.

LUCIA web site http://www.pd.istc.cnr.it/LUCIA/Docs/InterFace-AISV2004.pdf

Magno Caldognetto E. (2002). I correlati fonetici delle emozioni. In Bazzanella C., Kobau P. (Eds.), *Passioni, emozioni, affetti*. Milano: McGraw-Hill. 197-213.

Magno Caldognetto E., Ferrero F. E. (1996). Macro e micro variazioni fonetiche dipendenti dalle scelte paralinguistiche del parlante. In Fedi F., Paoloni A. (Eds.), Atti delle *VI Giornate di Studio del Gruppo di Fonetica Sperimentale - A.I.A. Caratterizzazione del parlante* (Roma, 23-24/ 11/ 1995). Roma: Esagrafica. 95-107.

Magno Caldognetto E., Poggi I. (2001). Dall'analisi della multimodalità quotidiana alla costruzione di Agenti Animati con Facce Parlanti ed Espressive. In Magno Caldognetto E., Cosi P. (Eds.), Atti delle *XI Giornate di Studio del GFS su Multimodalità e multimedialità nella comunicazione*. Padova: Unipress. 47-55.

Magno Caldognetto E., Vagges K. (1990). Il riconoscimento visivo dei movimenti articolatori da parte di soggetti normali e ipoacusici. In *Scritti in onore di Lucio Croatto*. Padova: Microprint. 153-166.

Magno Caldognetto E., Zmarich C.(1998). I visemi consonantici dell'italiano: la categorizzazione fonologica dei movimenti articolatori visibili. In Bertinetto P.B., Cioni L. (Eds.), Atti delle *VIII Giornate di Studio del G.S.F.*. Pisa, Italia. 192-202.

Magno Caldognetto E., Zmarich C. (1999). Visual Spatio-temporal Characteristics of Lip Movements in Defining Italian Consonantal Visemes. Proceedings of *ICPhS '99*. San Francisco, USA. Vol.2, 881-884.

Magno Caldognetto E., Zmarich C. (2001). L'intelligibilità dei movimenti articolatori visibili: caratteristiche degli stimoli vs. bias linguistici. In Magno Caldognetto E., Cosi P. (Eds.), Atti delle *XI Giornate di Studio del GFS su Multimodalità e multimedialità nella comunicazione*. Padova: Unipress. 33-40.

Magno Caldognetto E., Zmarich C., Cosi P., Ferrero F. (1997). Italian Consonantal Visemes: Relationships between Spatial/Temporal Articulatory Characteristics and Coproduced Acoustic Signal. Proceedings of the *Workshop on Audio-Visual Speech Processing Cognitive and Computational Approaches*. Rhodes, Greece. 5-8.

Magno Caldognetto E., Zmarich C., Ferrero F. (1998a). Indici acustici macro prosodici dello stato emotivo del parlante. Atti del *XXVI Convegno Nazionale di Acustica*, Torino. 263-266.

Magno Caldognetto E., Zmarich C., Cosi P. (1998b). Statistical Definition of Visual Information for Italian Vowels and Consonants. Proceedings of *Audio Visual Speech Processing '98*. Terrigal, AUS. 135-140.

Magno Caldognetto E., Cosi P., Cavicchio F. (2004 a). Modifications of Speech Articulatory Characteristics in the Emotive Speech. In André E., Dybkjaer L., Minker W., Heisterkamp P. (Eds.), Proceedings of the *Tutorial and Research Workshop*: *Affective Dialogue Systems*. Kloster Irsee, Germany. 233-239.

Magno Caldognetto E., Cosi P., Drioli C., Tisato G., Cavicchio F. (2004 b). Modifications of Phonetic Labial Targets in Emotive Speech: Effects of the Co-production of Speech and Emotions. *Speech Communication.* 44, 173-185.

Magno Caldognetto E., Poggi I., Cosi P., Cavicchio F., Merola G. (2004c). Multimodal Score: an ANVIL Based Annotation Scheme for Multimodal Audio-Video Analysis. Proceedings of the *Workshop on Multimodal Corpora Models of Human Behaviour for the Specification and*

Evaluation of Multimodal Input and Output Interfaces at the 4th International Conference on *L R E*. Lisbona, Portugal. 29-33.

Magno Caldognetto, Poggi I., Cosi I., Cavicchio F. (2005 a). Aspetti dell'interazione mediata da computer nell'e-learning: dall'analisi di chat e forum alla sintesi della Faccia Parlante. In Delfino M., Manca S., Persico D., Sarti L. (Eds.), *Come costruire conoscenza in rete?* Atti del III Workshop nazionale del progetto FIRB-MIUR "*Nuove tecnologie per la formazione permanente e reti nel sistema socioeconomico italiano*, Genova, 28 ottobre 2004. Genova: ITD CNR. 177-191.

Magno Caldognetto E., Cavicchio F., Cosi P., Drioli C., Tisato G. (2005 b). *Parametri per lo studio delle modificazioni articolatorie del parlato emotivo*. Atti *1°Convegno AISV*. Padova: EDK Press. 441-470.

Magno Caldognetto E., Cavicchio F., Poggi I. (in corso di stampa,a). L'espressione delle emozioni in chat, forum ed e-learning. In Magno Caldognetto E., Cavicchio F., Cosi P. (Eds.), Atti del 1° Convegno Nazionale GSCP (Gruppo di Studio della Comunicazione Parlata) su *Comunicazione Parlata e Manifestazione delle Emozioni*, Padova 30/10- 1/11 2004. Napoli: Liguori Editore.

Magno Caldognetto E., Cavicchio F., Cosi P. (in corso di stampa,b). La faccia e la voce delle emozioni. In Poggi I., Pascucci M. (Eds.), *La mente, gli altri e le emozioni*. Roma: Armando Editore.

Magno Caldognetto E., Cavicchio F., Cosi P. (in corso di stampa,c). Interfacce multimodali per l'e-learning. In Delogu C. (Ed.), *Tecnologia per il web learning: realtà e scenari.* Firenze: FUP Florence University Press.

Magno Caldognetto E., Cavicchio F., Cosi P., Poggi I. (in corso di stampa,d). Espressione delle emozioni e motivazione all'apprendimento: interfacce grafiche, vocali e multimodali a confronto. In Magno Caldognetto E., Cavicchio F. (Eds.), .), *Aspetti emotivi e relazionali nell'e-learning*. Firenze: FUP Florence University Press.

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

Massaro D.W. (1996). Bimodal Speech Perception: A Progress Report. In Stork D.G., Hennecke M.E. (Eds.*),Speechreading by Humans and Machines.* New York:: Springer-Verlag. 79-101.

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

Morishima Y., Nakajima H., Brave S., Yamada R., Maldonado H., Nass C., Kawaji S. (2004). The Role of Affect and Sociality in the Agent-Based Collaborative Learning System. In André E., L. Dybkjaer, W. Minker, P. Heisterkamp (Eds.), *Affective Dialogue Systems, Tutorial and Research Workshop, ADS 2004*. Berlin: Springer- – Verlag. 265-275.

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

MPEG-4 standard. Home page: http://www.chiariglione.org/mpeg/index.htm.

Nordstrand M., Svanfeldt G., Granstrom B., House D. (2004), Measurements of Articulatory Variation in Expressive Speech for a Set of Swedish Vowels. *Journal of Speech Communication.* 44, 187-196.

Pelachaud C. (2003). Emotion Expressiveness Embedded in Representation Languages for ECAs. Report written for the European Project pf-star http://pfstar.itc.it/public/doc/deliverables/pelachaud_tech_rep2.pdf

Pelachaud C., Badler N.I., Steedman M. (1996). Generating Facial Expression for Speech. *Cognitive Science*. 20, 1-46.

Picard R. (1998). *Affective Computing*. Cambridge (Mass.): The MIT Press.

Poggi I.,, Magno Caldognetto E. (1996). A Score for the Analysis of Gesture in Multimodal Communication. In L. Messing (Ed.), Proceedings of the *Workshop on the Integration of Gesture in Speech*. Newark and Wilmington, Delaware USA: Applied Science and Engineering Labs.. 235-244.

Poggi I., Magno Caldognetto E. (1998). A Procedure for the Generation of Gesture in Bimodal Communication. Proceedings of the *ORAGE '98 Colloque International Oralité et Gestualité* (Besancon 9-11/11/1998). Paris: L'Harmattan. 201-209.

Poggi I., Magno Caldognetto E. (2004). Il parlato emotivo. Aspetti cognitivi, linguistici e fonetici. In Albano Leoni F., Cutugno F., Pettorino M., Savy R. (Eds.), Atti del Convegno *Italiano parlato* (Napoli 14-15/2 2003). Napoli: D'Auria Editore, CD-Rom.

Poggi I., Pelachaud C. (2000). Performative Facial Expression in Animated Faces. In Cassell J., Sullivan J., Prevost S., Churchill E. (Eds.), *Embodied Conversational Agents*. Cambridge (Mass.): MIT Press. 155-188.

Poggi I., Pelachaud C., de Rosis F. (2000). Eye Communication in a Conversational 3D Synthetic Agent. AI Communications. 13, 169-181.

Poggi I., Pelachaud C., de Rosis F., Carofiglio V., De Carolis B. (2004). GRETA. A believable Embodied Conversational Agent. In Stock O., Zancanaro M. (Eds.), *Multimodal Intelligent Information Presentation*. New York: Kluwer. 1-23.

Poggi I., Bartolucci L., Violini S. (in corso di stampa). Emozioni. Un'arma per l'apprendimento. In Magno Caldognetto E., Cavicchio F. (Eds.), *Aspetti emotivi e relazionali nell'e-learning*. Firenze: FUP Florence University Press.

Spears R., Lea M., Postmes T. (2001). Social Psychological Theories of Computer-Mediated Communication: Social Pain or Social Gain. In Robinson W. P., Giles H. (Eds.), *The New Handbook of Language and Social Psychology*. Chichester (UK): J.Wiley & Sons. 601-624.

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

Tesser F., Cosi P., Drioli C., Tisato G. (2005). Emotional FESTIVAL-MBROLA TTS Synthesis. Proceedings of *INTERSPEECH 2005*. Lisbon, Portugal. CD-rom.

Tini Brunozzi F., Danieli M. (in corso di stampa). Da fonetico a grafemico. Manifestazione delle emozioni negli SMS tra oralità e scrittura. In Magno Caldognetto E., Cavicchio F. (Eds.), Atti del 1° Convegno Nazionale GSCP (Gruppo di Studio della Comunicazione Parlata) su *Comunicazione Parlata e Manifestazione delle Emozioni* (Padova 30/10-1/11 2004). Napoli: Liguori Editore.

Tisato G., Cosi P., Drioli C., Tesser F. (2004). INTERFACE: a New Tool for Building Emotive/Expressive Talking Heads. Atti 1° Convegno Nazionale *AISV 2004* su *Misura dei parametri*. Padova: EDK Editore. CD-rom.

Tisato G., Cosi P., Drioli C., Tesser F. (2005). INTERFACE: a Matlab© Tools for Building Animated MPEG4 Talking Heads from Motion-Captured Data. *Proceedings ICMI 2005*. Trento, Italia.

Ursini F. (2001). Multimodalità nella scrittura? Gli SMS tra telefoni cellulari, in Magno Caldognetto E., Cosi P. (Eds.), Atti delle *XI Giornate di Studio del Gruppo di Fonetica Sperimentale* su *Multimodalità e multimedialità nella Comunicazione* (Padova 29/11-1/12 2000). Padova: Unipress.75-80.

Gli autori

Emanuela Magno Caldognetto è Dirigente di ricerca CNR e responsabile della Sezione di Padova dell'ISTC. Ha dedicato le sue ricerche ai fenomeni di pianificazione ed esecuzione del parlato, studiando lapsus, pause, correlati acustici, articolatori e percettivi delle unità segmentali e soprasegmentali dell'italiano, e alla comunicazione multimodale, studiando la gestualità coverbale e la coproduzione di informazioni linguistiche e paralinguistiche nei segnali acustici e visivi che veicolano il parlato emotivo, anche in funzione della simulazione tramite Facce Parlanti ed Espressive e alla loro applicazione come interfacce uomo-macchina nell'e-learning. Ha diretto unità di ricerca e progetti nazionali, ha collaborato a progetti europei, è responsabile della Commessa "Parlato e Comunicazione Multimodale" del Dipartimento di Identità Culturale del CNR ed è docente di Linguistica e Fonetica in corsi di laurea, anche e-learning, della Facoltà di Medicina e Chirurgia dell'Università di Padova. E' stata coordinatrice nazionale del GFS (Gruppo di Fonetica Sperimentale A.I.A.), è membro del Comitato di coordinamento del GSCP, socia dell'AISV e membro di Humaine. E' autrice o coautrice di circa 200 tra articoli e libri.

Federica Cavicchio è attualmente PhD Student in Cognitive and Brain Sciences (CoBraS) presso il CIMeC (Centro Interdipartimentale Mente e Cervello) dell'Università di Trento, fa parte del CLIC Group (Cogniton, Language, Interaction and Computation) guidato dal Prof. Massimo Poesio. Ha lavorato come project manager e assistente di ricerca presso l'ISTC CNR di Padova per il progetto FIRB MIUR « Nuove Tecnologie e Reti nel Sistema Socioeconomico Italiano » sotto la responsabilità della Prof. Emanuela Magno Caldognetto. Ha avuto l'incarico di tutor on line negli anni accademici 2003/2006 del corso di «Glottologia e linguistica» per il corso di laurea in «Tecnico audiometrico e audioprotesico» della Facoltà di Medicina e Chirurgia dell'Università degli studi di Padova. E' membro della rete di eccellenza europea Humaine e coautore di numerose pubblicazioni.

Piero Cosi è Primo ricercatore presso l'ISTC CNR Sezione di Padova. La sua attività scientifica riguarda principalmente l'elaborazione e l'analisi del segnale verbale, lo studio dei modelli uditvi, il riconoscimento automatico del segnale verbale, la sintesi da testo scritto, lo studio dei sistemi basati sulle reti neurali, l'animazione facciale e gli Agenti parlanti. Più recentemente si è interessato al problema della sintesi audio/video del parlato emotivo ed espressivo. E' responsabile dello Speech and Multimodal Communication Laboratory presso cui vengono svolte attività di ricerca e sviluppo riguardanti le versioni italiane di alcuni software vocali: CSLU Speech Toolkit e CSLR SONIC per il riconoscimento automatico del segnale vocale; FESTIVAL e MBROLA per la sintesi da testo scritto; BALDI, GRETA e LUCIA per la creazione di Agenti espressivi ed emotivi. E' responsabile del progetto ILT (Italian Literacy Tutor), un programma integrato e interattivo per l'insegnamento della lingua italiana. E' autore o coautore di più di 100 articoli pubblicati su libri, riviste o atti di convegni internazionali. E' Presidente dell'AISV (Associazione Italiana di Scienze della Voce) e membro del Forum TAL (Trattamento Automatico del Linguaggio) del Ministero delle Poste e Telecomunicazioni. E'membro dell'ISCA e dell'IEEE.