FESTIVAL E LUCIA: TTS (Text-To-Speech) e IVA (Intelligent Virtual Agent) al servizio della didattica dei disabili

Piero Cosi, Emanuela Caldognetto Magno

ISTC-SPFD CNR
Istituto di Scienze e Tecnologie della Cognizione
Sede di Padova "Fonetica e Dialettologia"
Consiglio Nazionale delle Ricerche
Padova, Italy

web: www.pd.istc.cnr.it
e-mail: piero.cosi@pd.istc.cnr.it
e-mail: emanuela.magno@pd.istc.cnr.it

SOMMARIO

In questo lavoro vengono descritti *FESTIVAL* e *LUCIA,* due *tool* sviluppati presso l'Istituto di Scienze e Tecnologie della Cognizione - Sede di Padova "Fonetica e Dialettologia" del Consiglio Nazionale delle Ricerche di Padova, e alcune delle loro possibili applicazioni al servizio della didattica dei disabili.

FESTIVAL è un sintetizzatore vocale da testo scritto, espressivo ed emotivo. FESTIVAL, che è stato recentemente adattato all'italiano, è basato sulla tecnica della concatenazione di unità vocali. L'architettura generale del sistema comprende un blocco di Moduli Linguistici responsabili dell'*analisi testuale e linguistica* del testo in ingresso e da un blocco di Moduli Fonetico-Acustici responsabili dell'*analisi prosodica*, intesa come determinazione dell'intonazione e della durata, e della *generazione del segnale* che consente, quale ultimo passo, di generare una forma d'onda a partire dalle informazioni linguistiche sopra specificate.

LUCIA è un agente animato "*intelligente"* creato per migliorare l'interazione uomo-macchina e renderla più efficace e naturale. In altre parole, LUCIA è una faccia parlante in grado di esprimere espressioni ed emozioni ed il suo motore di animazione facciale è basato sullo standard MPEG-4.

INTRODUZIONE

Per superare le limitazioni della Comunicazione Mediata dal Computer (CMC) (Baracco, 2002) 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 che dovrebbero garantire maggiore naturalezza, intelligibilità e appropriatezza e risultare maggiormente persuasive.

Per creare tali Agenti, sulla base delle ricerche sulla generazione del linguaggio naturale e specificamente sul dialogo sviluppate soprattutto

nell'ambito dell'Intelligenza Artificiale 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 cognitiva delle conoscenze e delle intenzioni del Mittente, genera messaggi nelle diverse modalità utilizzate dagli umani nell'interazione faccia-a-faccia.

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 delle emozioni (Magno Caldognetto et al., in corso di stampa).

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.

FESTIVAL

FESTIVAL è un sistema di sintesi automatica emotiva ed espressiva da testo scritto[1] che, come già descritto in un precedente lavoro (Cosi et al., 2000), (Cosi et al., 2001), è stato recentemente realizzato anche per l'italiano. FESTIVAL è basato sulla tecnica della concatenazione di unità vocali.

L'architettura generale del sistema, illustrata schematicamente in Figura 1, comprende un blocco di Moduli Linguistici responsabili dell' "*analisi testuale e linguistica*" del testo in ingresso e da un blocco di Moduli Fonetico-Acustici responsabili dell' "*analisi prosodica*", intesa come determinazione dell'intonazione e della durata, e della "*generazione del segnale*" che consente, quale ultimo passo, di generare una forma d'onda a partire dalle informazioni linguistiche sopra specificate.

L'architettura dei Moduli Linguistici è illustrata in Figura 2. La stringa in ingresso è esaminata da un primo modulo che riconosce i dati numerici che

[1] http://www.cstr.ed.ac.uk/projects/festival/

sono direttamente trascritti a livello fonemico. I dati non numerici sono distinti a seconda che la parola sia una parola "*funzione*" o una parola "*contesto*". A questo punto le parole sono trascritte nella loro forma fonemica, o passando attraverso il lessico o applicando le regole esplicite di accentazione, trascrizione e sillabificazione. In particolare il modulo numerico espande a livello di parola e poi di fonemi i dati numerici distinguendo tra ore, date, numeri di telefono, ecc.

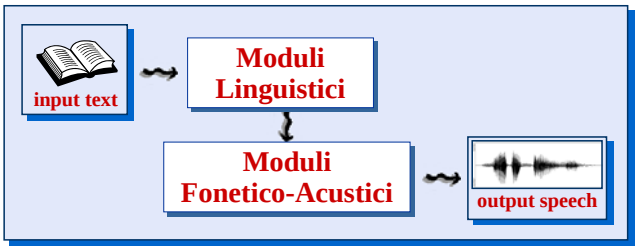

Figura 1. Architettura generale di un sintetizzatore da testo scritto.

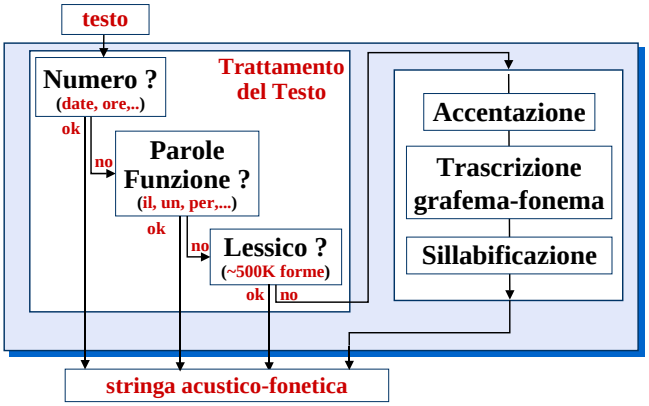

Figura 2. Architettura dei Moduli Linguistici per l'analisi del testo, l'accentazione, la trascrizione grafema-fonema e per la sillabificazione.

L'architettura dei Moduli Fonetico-Acustici è illustrata in Figura 3. A partire dalla stringa fonetica sin qui ottenuta sono selezionate le corrispondenti unità acustiche, nel nostro caso, i difoni, e per ognuna di esse è aggiunta l'informazione prosodica riguardante la durata e la frequenza fondamentale. Questi dati sono poi inviati al modulo di generazione vera e propria della forma d'onda che utilizza la sintesi LPC, eccitata dai residui ("*Residual Excited Linear Prediction",) o la sintesi MBROLA*[2].

[2] The MBROLA Project: http://tcts.fpms.ac.be/synthesis/

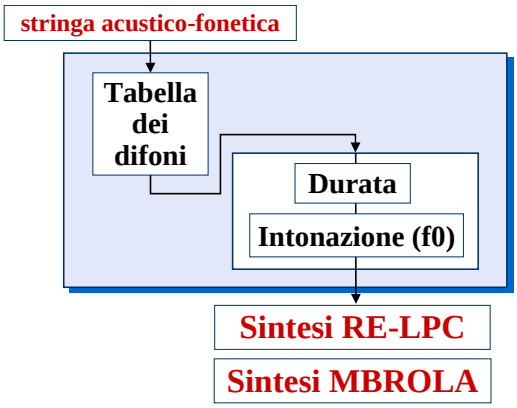

Figura 3. Architettura dei Moduli Fonetico-Acustici per l'assegnazione delle regole di durata e intonazione (f0) e per la generazione della forma d'onda mediante sintesi LPC o MBROLA.

LUCIA

LUCIA è una "faccia parlante" in italiano espressiva ed emotiva, sviluppata in questi ultimi anni (Cosi et al., 2003) presso i laboratori dell'ISTC-SPFD (Figura 4).

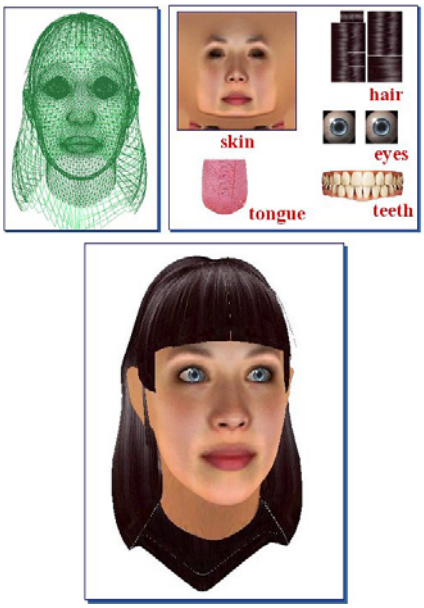

Figura 4. LUCIA: wireframe e texture.

Lucia è basata su uno specifico modello di coarticolazione (Cohen & Massaro, 1993) appositamente sviluppato per rendere più fluidi e naturali i movimenti delle labbra. LUCIA è basata sullo standard MPEG-4[3]. e parla in

[3] Mpeg-4 home page: http://www.chiariglione.org/mpeg/

italiano mediante la versione italiana di FESTIVAL (Cosi et al., 2000), (Cosi et al., 2001), la cui architettura è schematicamente illustrata nel diagramma a blocchi di Figura 5.

Il modello è visualizzato in tempo reale sullo schermo e sincronizzato con il corrispondente segnale vocale fornito dal sistema di sintesi da testo scritto. LUCIA è costruita mediante un reticolo di circa 27000 poligoni, ed il modello è diviso in due parti principali: la pelle e gli articolatori interni (occhi, lingua, denti). Questa suddivisione è particolarmente utile per l'animazione, poiché soltanto la pelle è direttamente influenzata dall'azione dei pseudo-muscoli mimici facciali è costituisce quindi un elemento unitario, mentre le altri componenti anatomiche risultano essere indipendenti fra loro e si muovono in modo più rigido seguendo esclusivamente delle traslazioni e/o rotazioni (per esempio gli occhi ruotano su se stessi attorno al loro punto centrale). Utilizzando questa strategia i poligoni sono distribuiti in modo tale da rendere l'effetto visivo molto naturale evitando di visualizzare possibili "discontinuità" nel modello 3D soprattutto in fase di animazione.

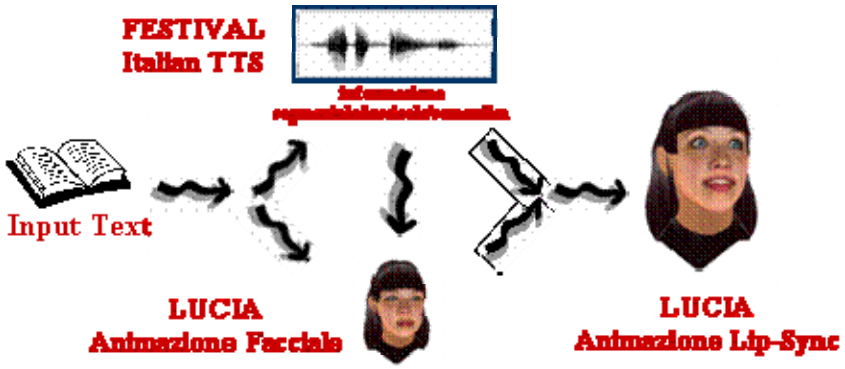

Figura 5 .Diagramma a blocchi dell'architettura di LUCIA.

In Figura 6 sono illustrate due espressioni di LUCIA tipiche di un eloquio normale e "spaventato".

Figura 6 . LUCIA "neutra" o "spaventata".

APPLICAZIONI

Gli agenti virtuali, come ad esempio LUCIA, attirano in modo particolare l'attenzione, soprattutto degli studenti e dei bambini, e producono di conseguenza una situazione di apprendimento particolarmente vivace ed efficace. E' proprio incorporando questi agenti animati espressivi ed emotivi in un ambiente di apprendimento multimodale e multimediale che ci si propone di migliorare le capacità di lettura degli allievi mediante gli esercizi sulle abilità di base e mediante i libri interattivi.

LUCIA, infatti, è inserita con queste finalità in un progetto, per ora realizzato solo in parte sotto forma di prototipo, avente come obiettivo principale la creazione di un sistema di apprendimento linguistico per l'Italiano, come L1 o L2, all'interno del quale vengono sviluppati inoltre, con tecnologie avanzate, utili strumenti automatici per il feedback.

In questo ambiente di sviluppo LUCIA funge da intermediario o da interfaccia uomo-macchina, ed è munita di espressioni e gesti del viso adeguati a ciascun compito. Inoltre LUCIA si esprime oralmente ed utilizza il riconoscimento automatico della voce per interagire con l'utente studente.

Di conseguenza, lo sviluppo di questo progetto di natura prevalentemente applicativa, porterà un sensibile avanzamento della ricerca di base in molteplici settori quali l'animazione facciale, la sintesi automatica da testo scritto, il riconoscimento automatico del parlato e l'analisi linguistica completa con strumenti sintattici "robusti".

Vi sono inoltre molte altre possibili applicazioni in cui una faccia parlante potrebbe essere efficacemente utilizzata quali ad esempio le CHAT, le e-mail animate, la lettura animata di fiabe per bambini o in generale di libri interattivi, le presentazioni o le lezioni animate, i giochi al computer, solo per citarne alcune, ed in generale tutte quelle applicazioni in cui una presenza visuale attiva può essere di aiuto per focalizzare l'attenzione o per vivacizzare l'apprendimento.

OSSERVAZIONI CONCLUSIVE

In conclusione possiamo sottolineare come l' e-learning[4] pur essendo basato su una forte base tecnologica, diventata ormai altamente affidabile, sia anche contemporaneamente orientato su una forte base pedagogica, essendo, infatti, un processo fondamentalmente sociale che dovrebbe facilitare l'interazione e la collaborazione fra le persone.

L'e-learning sta sicuramente rivoluzionando, in termini di organizzazione, il rapporto docente-discente ed in generale tutte le attuali teorie sull'insegnamento e l'apprendimento e le "Facce Parlanti" possono

[4] e-learning europa, www Page. http://www.elearningeuropa.info/

certamente inserirsi in un tale scenario contribuendo a realizzare sistemi di insegnamento/apprendimento interattivi, vivaci ed efficaci.

WWW

FESTIVAL
http://www.pd.istc.cnr.it/FESTIVAL

LUCIA
http://www.pd.istc.cnr.it/LUCIA

BIBLIOGRAFIA

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.

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

Cosi P., Gretter R., Tesser F., (2000), "FESTIVAL parla italiano!". *Atti XI Giornate di Studio del G.F.S.*, Padova, Italy, November 29-30, Dicember 1, 2000, 235-242.

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

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

Magno Caldognetto E., Cavicchio F., Cosi P. (in corso di stampa, a). La faccia e la voce delle emozioni. In Poggi I. (Ed.), La mente del cuore.Scienze cognitive ed emozioni. Roma: Armando Editore.