“BALDI”…. PARLERA' ITALIANO?

Piero Cosi

IFD-CNR
Istituto di Fonetica e Dialettologia – Consiglio Nazionale delle Ricerche
e-mail: cosi@csrf.pd.cnr.it www: http://nts.csrf.pd.cnr.it/Ifd

1. SOMMARIO

Dopo il riconoscimento automatico e la sintesi da testo scritto, per ultimare l’*italianizzazione* del sistema denominato CSLU Speech Toolkit è necessario far parlare BALDI in Italiano. In questo lavoro sono descritte le principali linee guida per lo sviluppo del sistema e vengono illustrati alcuni esempi di animazione facciale realizzati mediante l’utilizzo simultaneo di BALDI e di FESTIVAL per l’italiano[1].

2. INTRODUZIONE

Lo sviluppo di nuove tecnologie, nel campo del trattamento automatico del linguaggio (TAL), è d'enorme interesse per un'efficace utilizzazione di nuove strumentazioni multi-mediali rivolte alla semplificazione dell'interfaccia uomo-macchina e le ricadute scientifico/applicative di queste tecnologie riguardano una serie innumerevole di applicazioni. Negli ultimi anni la riproduzione di un testo scritto tramite una voce sintetica (Text-To-Speech, TTS) ha raggiunto un grado di qualità tale da poter essere sfruttato proficuamente in moltissime applicazioni [1]. Allo stesso, modo nel campo della grafica digitale si sono fatti progressi, un tempo inimmaginabili [2], che hanno consentito la realizzazione di veri e propri Agenti Virtuali in grado di interagire con l’utente, nel caso ad esempio di sistemi di comunicazione uomo-macchina, in modo semplice e naturale [3-4]. La naturalezza della “voce” sintetica come pure la verosimiglianza delle espressioni facciali ad essa collegate sono ovviamente il punto focale di questi sistemi e ne determinano la loro efficace applicabilità ed utilizzazione.

Relativamente alla sintesi audio, già da molto tempo, esistono sistemi in grado di “parlare” in modo intelligibile, in cui l'ascoltatore, cioè, riesce a capire il contenuto del messaggio prodotto, a distinguere le parole e a individuarne il significato. In questi sistemi però è indiscutibile riconoscere l’identità sintetica del parlante, perché alcune delle caratteristiche naturali del linguaggio parlato umano, quali ad esempio la corretta coarticolazione della pronuncia dei fonemi, l'intonazione, il ritmo, la fluidità e altri parametri ancora, non sono correttamente simulati. La ricerca in questo settore, grazie soprattutto allo sviluppo di nuovi modelli teorici e di nuove tecniche digitali di sintesi, alla simultanea e vertiginosa crescita delle capacità computazionali dei calcolatori e alla facile

[1] Parte di questo lavoro è stato svolto nell'ambito dei progetti di ricerca denominati MPIRO (Multilingual Personalized Information Objects. European Project IST-1999-10982 - WWW page: http://www.ltg.ed.ac.uk/mpiro/), finanziato dalla Comunità Europea, e TICCA (Tecnologie cognitive per l'interazione e la cooperazione con agenti artificiali), finanziato dal CNR e dalla Provincia Autonoma Trentina.

reperibilità di enormi librerie di dati vocali, ha affinato le tecniche di sintesi portando a risultati più che soddisfacenti in ormai quasi tutte le principali lingue.

Allo stesso modo, le tecniche digitali di animazione sono ormai arrivate a livelli talmente sofisticati per cui è spesso difficile distinguere gli attori "umani" da quelli "sintetici", cioè animati artificialmente, nelle produzioni cinematografiche o nei programmi interattivi e nei giochi al computer Pur tuttavia rimane da sottolineare che la riproduzione artificiale di espressioni umane è comunque ben diversa dalla loro generazione automatica a partire da un testo scritto e per questa devono continuamente essere sviluppate ulteriori nuove ricerche e nuovi modelli.

Come recentemente testimoniato dai Report CNR 2000 e 2001 [5], le attività di ricerca dell'IFD sono ormai da alcuni anni all'avanguardia in questo settore, e in particolare, per quanto riguarda la multimedialità, sono rivolte essenzialmente alla creazione di nuovi "agenti parlanti" ("talking agent"), virtuali ed interattivi, in grado di fornire all'utente un adeguato "feed-back" audio-visivo al fine di garantire la naturalezza e la semplicità di utilizzo di sistemi di interfaccia *uomo-macchina,*, in grado cioè di "*comunicare in modo naturale*" con l'utente.

3. "BALDINI"

In collaborazione con il *Perceptual Science Laboratory* (PSL) della *University of California, Santa Cruz* (UCSC) diretto dal Prof. Dominique Massaro è ormai da alcuni anni che stiamo lavorando alla creazione di BALDINI, un agente parlante in italiano.

BALDINI si basa interamente su un sofisticato modello di animazione facciale denominato BALDI sviluppato da Michael Cohen e Dominique Massaro del PSL per l'inglese [6]. In Figura 1 sono illustrati alcuni esempi di BALDI, mentre in Figura 2 è descritto lo schema generale utilizzato per la realizzazione di BALDINI.

Per quanto riguarda la voce di BALDINI è stata utilizzata la voce italiana di FESTIVAL [7], sviluppata dall'IFD in collaborazione con l' "Istituto Trentino di Cultura - Centro per la Ricerca Scientifica e Tecnologica (ITC-IRST) di Trento, i cui più recenti sviluppi sono descritti in un recente lavoro in questo stesso volume [8].

Per la realizzazione di BALDINI è innanzi tutto necessario identificare e categorizzare tutti i "*visemi*" specifici per la lingua italiana (la configurazione visiva delle labbra, dei denti, della lingua, della mandibola, ecc.) durante la produzione dei corrispondenti fonemi ed è a questo scopo che sono state e continueranno ad essere orientate numerose ricerche dell'IFD [9-11].

Modulo essenziale del sistema è poi quello che consente una corretta sincronizzazione fra il segnale verbale, prodotto dal sistema di sintesi da testo scritto, e i movimenti articolatori coinvolti nella sua produzione, generati dal "motore" di animazione facciale BALDI.

In questa prima fase della ricerca i movimenti articolatori di BALDINI sono i movimenti corrispondenti ai visemi inglesi che sono stati "mappati" sui corrispondenti visemi italiani, come risulta dalla Tabella 1. In futuro, i movimenti articolatori di BALDINI, in particolare i movimenti labiali, saranno appresi sugli effettivi dati analitici raccolti per l'italiano in questi ultimi anni e la naturalezza del sistema potrà ulteriormente migliorare.

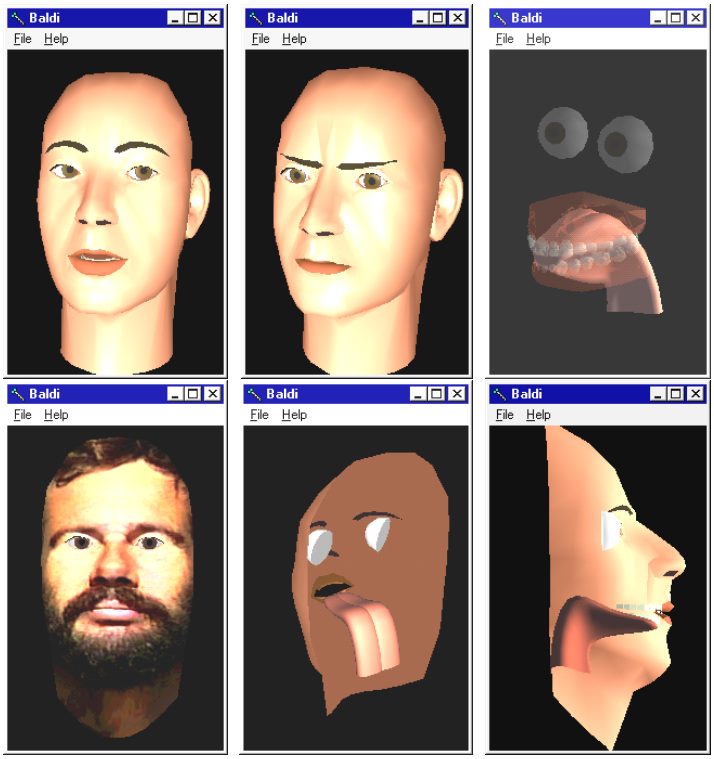

Figura 1. BALDI. Un sofisticato modello per l'animazione facciale.

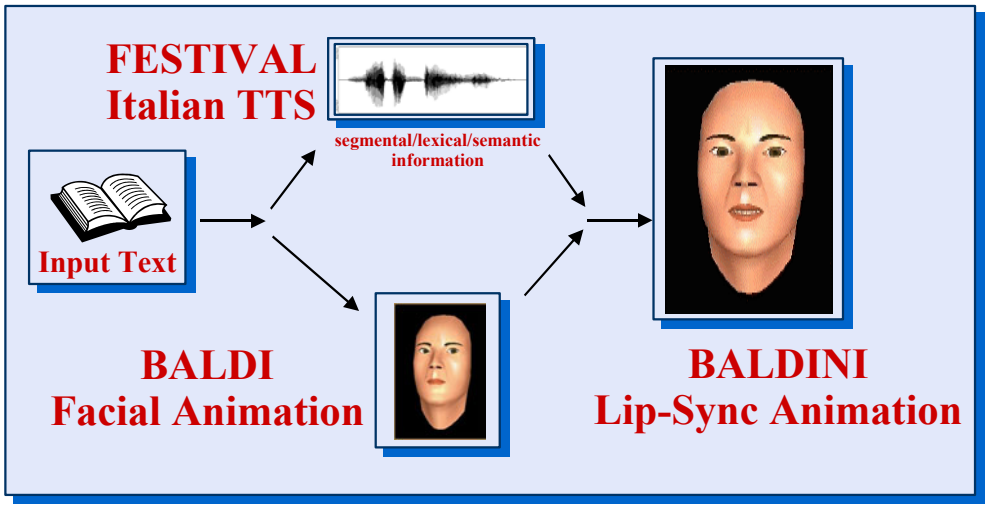

Figura 2. Schema generale utilizzato per la realizzazione di "BALDINI".

p	pane	bread	"**p**ane	**p**	**pin**	**pIn**	**p**
b	bacio	kiss	"**b**atSo	**b**	**bin**	**bIn**	**b**
t	torre	tower	"**t**orre	**t**	**tin**	**tIn**	**t**
d	danno	damage	"**d**anno	**d**	**din**	**dIn**	**d**
k	cane	dog	"**k**ane	**k**	**kin**	**kIn**	**k**
g	gamba	leg	"**g**amba	**g**	**give**	**gIv**	**g**
ts	zitto	silent	"**ts**itto	**t + s**	**pet sin**	**pet sin**	**t+s**
dz	zona	zone	"**dz**Ona	**d + z**	**pod zing**	**pQd zIN**	**d+z**
tS	cena	dinner	"**tS**ena	**tS**	**chin**	**tSIn**	**ch**
dZ	gita	outing	"**dZ**ita	**dZ**	**gin**	**dZIn**	**jh**
f	fame	hunger	"**f**ame	**f**	**fin**	**fIn**	**f**
v	vano	vain	"**v**ano	**v**	**vim**	**vIm**	**v**
s	sano	healthily	"**s**ano	**s**	**sin**	**sIn**	**s**
z	sbaglio	mistaken	"**z**baLLo	**z** (solo in cluster iniziali)	**zing**	**zIN**	**z**
S	scena	scene	"**S**Ena	**S**	**shin**	**SIn**	**sh**
m	molla	spring	"**m**Olla	**m**	**mock**	**mQk**	**m**
n	nave	ship	"**n**ave	**n**	**knock**	**nQk**	**n**
J	gnocco	lump	"**J**Okko	**dZ**	**gin**	**dZIn**	**jh**
r	rete	network	"**r**ete **l**	**long**	**lQN**	**l**	
l	lama	blade	"**l**ama	**l**	**long**	**lQN**	**l**
L	gli	the (plural)	**L**i (solo in *function words*)	**dZ**	**gin**	**dZIn**	**jh**
j	ieri	yesterday	"**j**Eri **j**	**yacht**	**jQt**	**y**	
w	uomo	man	"**w**Omo	**w**	**wasp**	**wQsp**	**w**
i	vita	life	"v**i**ta **I**	**vim**	**vIm**	**ih**	
e	rete	network	"r**e**te **EI**	**h**	**EI ch**	**ih**	
E	meta	goal	"m**E**ta	**E**	**bet**	b**E**t	**eh**
a	rata	rate	"r**a**ta **V**	**cut**	**kVt**	**ah**	
O	moto	motion	"m**O**to	**A**	**pot**	**pAt**	**aa**
o	dove	where	"d**o**ve	**@U**	**over**	**@Uvr=**	**ow**
u	muto	dumb	"m**u**to	**U**	**put**	**pUt**	**uh**

Tabella 1. Mappatura utilizzata per far corrispondere i fonemi inglesi a quelli italiani.

4. APPLICAZIONI

Le possibili applicazioni di facce parlanti espressive e naturali sono intuitivamente numerosissime e vanno dai servizi di telecomunicazioni, ai sistemi per l'insegnamento e l' apprendimento o per l'aiuto alle persone disabili, ai libri e giocattoli animati e parlanti, ai sistemi di comunicazione uomo-macchina, ai sistemi multimediali

L'introduzione degli agenti virtuali parlanti emotivi, abbinati ad un opportuno sistema computerizzato potrà contribuire sicuramente, ad esempio, a rendere più efficace ed interattiva l'utilizzazione, soprattutto da parte di studenti in età prescolare e scolare, di

sistemi automatici per l'insegnamento e l'apprendimento, cone nell'esempio illustrato in Figura 3.

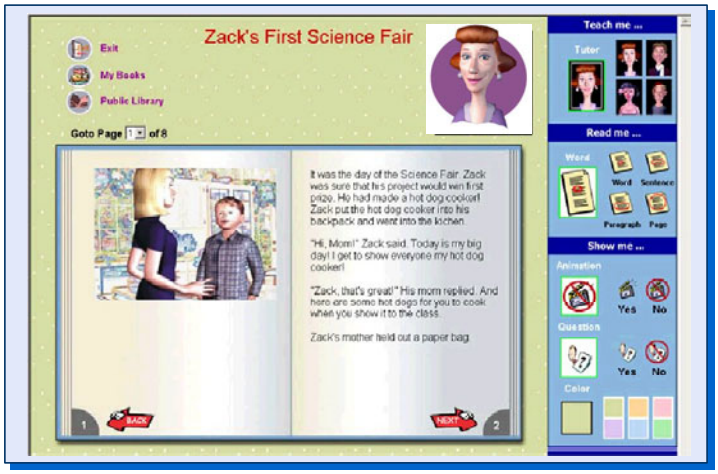

Figura 3. Esempio di "Libro Parlante" sviluppato al "Center for Spoken Language Research" (CSLR) della "Colorado University" (CU) di Boulder, CO, USA.

Nel campo della disabilità, come già accennato, i disabili visivi possono trarre enormi vantaggi da un sistema di riproduzione vocale naturale, ad esempio, mediante la semplice fruizione di materiale scritto o l'interazione con i sistemi operativi dei PC in tutti quei lavori, sempre più diffusi, che richiedano l'uso assiduo dei calcolatori. Introducendo inoltre la modalità visiva nella comunicazione uomo-macchina gli agenti virtuali parlanti possono consentire alle persone che soffrono di disturbi uditivi di poter apprendere più efficacemente le caratteristiche del linguaggio parlato (vedi Figura 4).

Figura 4. Bambini della Tuker-Maxon Oral School di Portland Oregon durante le loro lezioni sull'apprendimento del linguaggio.

4. CONCLUSIONI E SVILUPPI FUTURI

Una volta ultimato, BALDINI sarà integrato nell'ambiente software CSLU Speech Toolkit [12] sviluppato per l'inglese dal "Center for Spoken Language Understanding" (CSLU) dell' "Oregon Graduate Institute" (OGI) di Portland con cui collaboriamo ormai da molti anni.

Si può certamente affermare che almeno nel breve periodo, gli agenti virtuali certo non sostituiranno gli insegnanti umani, ma sicuramente li affiancheranno con frequenza e diffusione sempre maggiori soprattutto al fine di rendere più efficaci le lezioni e più produttivo l'apprendimento.

Una delle principali novità che sono e saranno sempre più introdotte mediante l'utilizzazione di agenti virtuali parlanti emotivi sarà la possibilità di trasmettere non solo il significato del messaggio, ma anche le emozioni e i possibili sentimenti ad esso collegati. Questa caratteristica rende la multimedialità nella comunicazione un passo fondamentale e necessario per una sempre maggiore naturale interazione uomo-macchina e le recenti apparizioni di insegnanti, giornalisti ed annunciatori/trici virtuali in internet ed in sistemi di apprendimento ed intrattenimento ne sono una prova significativa.

BIBLIOGRAFIA

[1] Van Santen et al. (editors), *Progress in Speech Synthesis.* Springer Verlag New York, Inc. 1997.

[2] D.W. Massaro, ***Perceiving Talking Faces****. From Speech Perception to a Behavioral Principle.* MIT Press, 1998.

[3] Cole R., Hirschman L., Atlas L. et al., The Challenge of Spoken Language Systems: Research Directions for the Nineties. *IEEE Trans. on Speech and Audio Processing*, Vol. 3, No. 1, January 1995. pp. 1-21.

[4] Cole R., Carmell T., Connors P., Macon M., Wouters J., de Villiers J., Tarachow A., Massaro D., Cohen M., Beskow J., Yang J., Meier U., Waibel A., Stone P., Fortier G., Davis A., Soland C., Intelligent Animated Agents for Interactive Language Training. In *Proceedings of STiLL (ESCA Workshop) Speech Technology in Language Learning*, Marholmen, Sweden, May 1997.

[5] *La comunicazione umana cede il passo a quella virtuale.* CNR Report 2001. http://www.presidenza.cnr.it/report2001/pdf/164166fo.pdf.

[6] *Sistemi automatici per l'interazione uomo-macchina*. CNR Report 2000. http://www.presidenza.cnr.it/report2000/pdf/focus/focus13.pdf.

[7] Massaro D.W., Cohen M.M., Beskow J., Cole R.A., Developing and evaluating conversational agents. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors, *Embodied Conversational Agents*, MIT Press, Cambridge, MA, 2000, pp. 287-318.

[8] FESTIVAL. A.W. Black (awb@cs.cmu.edu), P. Taylor (Paul.Taylor@ed.ac.uk), R Caley, R. Clark (robert@cstr.ed.ac.uk), CSTR - Centre for Speech Technology - University of Edinburgh. WWW page: http://www.cstr.ed.ac.uk/projects/festival/.

[9] Cosi P., Gretter R., Tesser F., Recenti sviluppi di FESTIVAL per l'italiano. In *Atti XII Giornate di Studio del G.F.S.*, Italy, November 29-30, Dicember 1, 2001, (in questo volume).

[10] Cosi P., Magno Caldognetto E., Lip and Jaw Movements for Vowels and Consonants: Spatio-Temporal Characteristics and Bimodal Recognition Applications. In *Speechreading by Humans and Machine: Models, Systems and Applications*, D.G. Storke and M.E. Henneke eds., NATO ASI Series, Series F: Computer and Systems Sciences, Vol. 150, Springer-Verlag, 1996, pp. 291-313.

[11] Magno Caldognetto E., Zmarich C., Cosi P., Statistical definition of visual information for Italian vowels and consonants. In Burnham D., Robert-Ribes J., Vatikiotis-Bateson E.(Eds), *Proceedings of the International Conference on Auditory-Visual Speech Processing, AVSP'98*,Terrigal, (AUS). pp. 135-140.

[12] Magno-Caldognetto E., Zmarich C., Visual Spatio-Temporal Characteristics of Lip Movements in Defining Italian Consonantal Visemes. In *Proceedings of ICPhS '99*, San Francisco, California (USA), vol 2, pp. 881-884.

[13] Sutton S., Cole R., Villiers J., Schalkwyk J., Vermeulen P., Macon M., Yan Y., Kaiser E., Rundle B., Shobaki K., Hosom P., Kain A., Wouters J., Massaro D., Cohen M., Universal speech tools: the CSLU toolkit. In Proceedings of ICSLP-98, Sydney, Nov 30-Dec 4, 1998, Vol. 7, pp. 3221-3224.