Cosi, P., Cohen, M.M. & Massaro, D.W. (2002). Baldini: Baldi speaks Italian. In proceedings of *7th International Conference on Spoken Language Processing, (ICSLP'02)* (pp.2349-2352). Denver, CO.

BALDINI: BALDI SPEAKS ITALIAN!

Piero Cosi, Michael M. Cohen** and Dominic W. Massaro***

*ISTC - Istituto di Scienze e Tecnologie della Cognizione
SFD - Sezione di Fonetica e Dialettologia, CNR - Consiglio Nazionale delle Ricerche
e-mail: cosi@csrf.pd.cnr.it WWW URL: http://nts.csrf.pd.cnr.it/

**Department of Psychology - Social Sciences II
University of California-Santa Cruz, Santa Cruz, CA 95064 USA
e-mail: mmcohen@ranx.ucsc.edu WWW URL: http://mambo.ucsc.edu/psl
e-mail: massaro@fuzzy.ucsc.edu

ABSTRACT

In this work, the development of Baldini, an Italian version of Baldi, a computer-animated conversational agent, is presented. Speech synthesis and facial animation examples are shown[1].

1. INTRODUCTION

The development of new technologies in the field of the speech and language science is essential for an effective utilization of new multimodal and multimedia systems aimed to simplify man-machine interfaces, enabling people to communicate with machines using natural communication skills. Advances in human language technology offer the promise of user-friendly and nearly universal access to on-line information and services. Since almost everyone speaks and understands a language, the development of spoken language systems will allow the average person to interact with computers without special skills or training, using common devices such as the telephone or personal computer.

The human face presents visual information during speech production that is critically important for effective communication. While the voice alone is usually adequate for communication, visual information from movements of the lips, tongue and jaw enhance intelligibility of the message (as is readily apparent with degraded auditory speech). For individuals with severe or profound hearing loss, understanding visible speech can make the difference between communicating effectively with others or a life of relative isolation. Moreover, speech communication is further enriched by the speaker's facial expressions, emotions, and gestures [1].

One goal of the Perceptual Science Laboratory (PSL) has been to create animated agents that produce accurate auditory and visible speech, as well as realistic facial expressions, emotions, and gestures. The invention of such agents has a tremendous potential to benefit virtually all individuals, but especially those with hearing problems, including the millions of people who acquire age-related hearing loss every year, and for whom visible speech and facial expression take on increasing importance. The animated characters that we are developing can be used to train individuals with hearing loss to "read" visible speech and to improve their processing of limited auditory speech, and will thereby facilitate face-to-face communication and access to online information presented orally by either real or lifelike computer characters.

The conversational agent, Baldi, is a 3-D animated talking head appropriately aligned with either synthesized or natural speech. The quality and intelligibility of Baldi has been repeatedly evaluated and shown to accurately simulate much of the behavior of naturally talking humans [1]. Baldi has a tongue and palate, which have been modeled on real tongues and palates. To illustrate these internal articulators, Baldi can display a midsagital view, or the skin on the face can be made transparent. These features allow one to pedagogically illustrate correct articulations and reveal the internal structures of the mouth usually hidden by the face. We are currently evaluating a motivational language tutorial program capable of introducing classroom curriculum, including training and developing language and listening skills [2-3].

The goal of the current work is to have Baldi speak Italian. When he does so, we call him Baldini. In the last years the reading of written texts by synthetic voices (Text-To-Speech synthesis, TtS) has achieved an improved quality and has been introduced and effectively applied in numerous applications [4]. In the same way, the field of facial animation has advanced to allow the introduction of Virtual Agents [5], able to naturally interact with the user, in man-machine communication systems [6-7]. The "natural" quality of the synthetic voice and the human likeness of a facial animated character are reasonable focal points that most likely determine their effective usability. As for TtS, segment intelligibility is mostly a solved problem, but the systems still maintain a noticeable synthetic quality, mostly because some of the features characterizing human speech, such as for example the correct coarticulation among phonemes, the intonation, the rhythm, the spontaneous fluency and so on, are not so well simulated by the current synthetic

[1] Part of this work has been sponsored by MPIRO (Multilingual Personalized Information Objects. European Project IST-1999-10982 - WWW page: http://www.ltg.ed.ac.uk/mpiro/) and TICCA (Tecnologie cognitive per l'interazione e la cooperazione con agenti artificiali) a joint "CNR-Provincia Autonoma Trentina" Project. The research at PSL was supported by grants from the National Science Foundation (Grant No. CDA-9726363, Grant No. BCS-9905176, Grant No. IIS-0086107), Public Health Service (Grant No. PHS R01 DC00236)

TtS synthesis engines. However, advances in this field, mostly thanks to the development of new theoretical models, new digital techniques, and the incredible increase of computational power and storage capability of computers have been so rapid during the last decade that good and quite natural TtS systems are beginning to appear not only in the scientific literature but also in the marketplace. At the same time, digital animation techniques have improved so dramatically that sometimes it is difficult to distinguish between human and synthetic actors in movie productions or in interactive games. However the automatic generation of human like speech and expressions driven directly from a written text is still far from a solved problem task and new advanced models and techniques will have to be developed in the future.

ISTC-SFD is deeply involved in the field of these activities, especially in multi-media research oriented to the creation of virtual and interactive expressive "Talking Agents" able to "communicate" with the user in human-machine communication systems [8-9].

2. BALDINI

A new Italian Talking Agent named Baldini has been recently developed, and soon it will be tested in perceptual evaluation tests. Baldini is a recent adaptation of the Baldi facial animation engine to the Italian language, whose architecture block diagram is introduced in Figure 1.

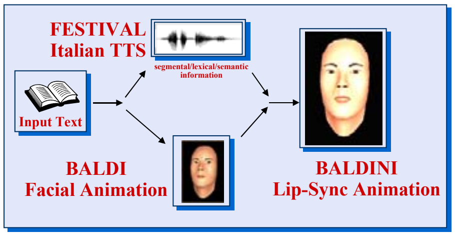

Figure 1: General architecture of Baldini.

Baldini's voice is the Italian version of FESTIVAL [13], developed with FESTVOX tool [14] by ISTC-SFD in cooperation with ITC-IRST ("Istituto Trentino di Cultura - Centro per la Ricerca Scientifica e Tecnologica) in Trento (Italy), whose most recent developments are described in a paper in this volume [15]. Baldini's visual speech is based on previous studies of Italian specific "visemes" (the visual configuration of lips, teeth, tongue and jaw during production of phonemes) [10-12], and is aligned with his Italian voice.

2.1. FESTIVAL

Festival is a general multi-lingual speech synthesis system developed at the CSTR (Center for Speech Technology Research, Edinburgh, Scotland, UK) [13] offering a full text to speech system with various APIs (Application Program Interfaces), as well an environment for development and research of speech synthesis techniques. It is written in C++ with a SCHEME-based command interpreter [16] for general control. Festival is a general-purpose concatenative text-to-speech (TtS) system that uses the residual-LPC and MBROLA [17] synthesis techniques, and is able to transcribe unrestricted text to speech. Festival contains standard tricks for intonation and duration prediction. Generally intonation is generated in two steps: prediction of accents and prediction of F0. In the simplest case, the intonation parameter is just set to constant values in the start and at the end of the utterance (130, 110 Hz), and more complex modules use so called Classification and Regression trees (CART). The Italian Festival modules [18] are described in Figures 2 and 3.

A first module implements a simple grammar to convert strings of numbers into word sequences. Numbers are expanded at the word and phoneme level distinguishing among time, dates, telephone numbers, etc. Alphabetic text is divided into "function-words" and "content-words". In the prosodic modules, "function words" are treated differently from the "content-words". The "content-words" are phonemically transcribed by the use of a wide lexicon, compiled in Festival format to speed up search procedures. If they are not present in the lexicon, they are phonemically transcribed by the use of explicit stress-assignment, letter-to-sound and syllabification rules. The Lexicon, compiled in Festival format, comprises 500k stressed Italian phonemically transcribed forms divided in syllables and labeled with their grammatical class or part-of-speech (POS). The correct diphones are then extracted from the acoustic database, where for each unit, specific information relative to its mean duration and pitch is included, and, finally, duration and intonation (F0) assignment is applied before activating the speech synthesis engine (residual LPC, Mbrola) to generate the speech waveform.

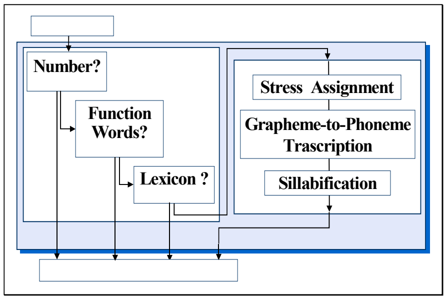

Figure 2: Text/Linguistic module for Italian FESTIVAL.

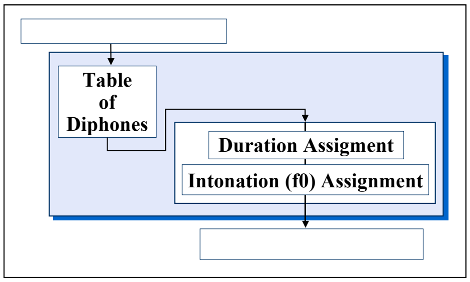

Figure 3: Prosodic module for Italian FESTIVAL *followed by the synthesis engine module (RE_LPC and MBROLA)*

2.2. BALDINI

To create the first baseline version of Baldini, correspondences between English and Italian phonemes were noted and implemented. In other words, at this stage of the implementation, the visible renditions of the English phonemes were mapped to the corresponding visible renditions of the Italian phonemes. Figure 4 presents some possible Baldini configurations. Table 1 gives the mapping between English and

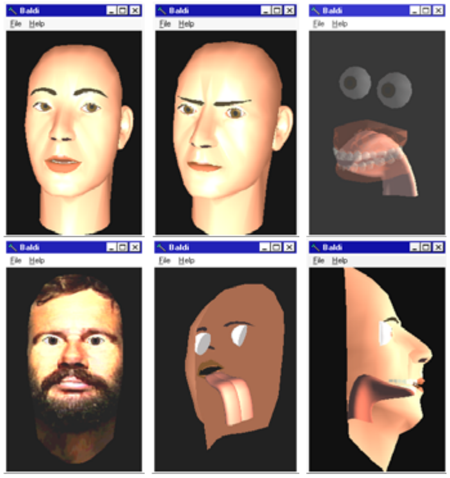

Figure 4:Some possible Baldini configurations.

Italian phonemes, indicated with SAMPA [19] and OGI [20] symbols, together with reference words for both languages.

We plan to first evaluate this implementation and then improve the observed limitations in Baldini's articulation by analyzing Italian articulatory movements from a big Italian corpus.

Baldini will be driven by either the TtS system or by naturally recorded speech in Italian and, in the latter case, the system designer need only type in the text of the recorded speech, and the system will transcribe the speech phonetically for synchronization with the talking face. Even if, English and Italian phonemes/visemes are obviously different in their acoustic and visual realization, and also few Italian phonemes, such as for example J in "gnocco" or L in the article "gli", are not present in English and have to be substituted with similar ones, the overall quality of Baldini looks quite acceptable at a first audio and visual inspection, but much more complete and accurate perceptual experiments have to be designed in the future to scientifically support these observations. When Baldini is trained on real Italian articulatory movements on a big Italian corpus, we can be confident that the quality of the system will increase in terms of naturalness and accuracy. Finally, the Italianization of the CSLU Speech Toolkit [21] will have automatic speech recognition [22-23] and text to speech synthesis [15], along with Baldini. These speech applications for Italian will be freely available for research purposes.

SAMPA	**Italian**	**English**	**SAMPA**	**English**	**English**		**OGI**
symbol	**word**	**translation**	**Transcription**	**phoneme**	**word**	**transcription**	
p	pane	bread	"pane	p	pin	pIn	p
b	bacio	kiss	"batSo	b	bin	bIn	b
t	torre	tower	"torre	t	tin	tIn	t
d	danno	damage	"danno	d	din	dIn	d
k	cane	dog	"kane	k	kin	kIn	k
g	gamba	leg	"gamba	g	give	gIv	g
p	pane	bread	"pane	p	pin	pIn	p
ts	zitto	silent	"tsitto	t + s	pet sin	pet sin	t+s
dz	zona	zone	"dzOna	d + z	pod zing	pQd zIN	d+z
tS	cena	dinner	"tSena	tS	chin	tSIn	ch
dZ	gita	outing	"dZita	dZ	gin	dZIn	jh
f	fame	hunger	"fame	f	Fin	fIn	f
v	vano	vain	"vano	v	vim	vIm	v
s	sano	healthily	"sano	s	Sin	sIn	s
z	sbaglio	mistaken	"zbaLLo	z	zing	zIN	z
			(solo in cluster iniziali)				
S	scena	scene	"SEna	S	shin	SIn	sh
m	molla	spring	"mOlla	m	mock	mQk	m
n	nave	ship	"nave	n	knock	nQk	n
J	gnocco	lump	"JOkko	dZ	gin	dZIn	jh
r	rete	network	"rete	l	long	lQN	l
l	lama	blade	"lama	l	long	lQN	l
L	gli	the (plural)	Li	dZ	gin	dZIn	jh
			(solo in function words)				
j	ieri	yesterday	"jEri	j	yacht	jQt	y
w	uomo	man	"wOmo	w	wasp	wQsp	w
i	vita	life	"vita	I	vim	vIm	ih
e	rete	network	"rete	EI	h	EI ch	ih
E	meta	goal	"mEta	E	bet	bEt	eh
a	rata	rate	"rata	V	cut	kVt	ah
O	moto	motion	"mOto	A	pot	pAt	aa
o	dove	where	"dove	@U	over	@Uvr=	ow
u	muto	dumb	"muto	U	put	pUt	uh

Table1: English to Italian Phoneme/Viseme mapping.

3. CONCLUDING REMARKS

Character animation with realistic speech and emotion offers dramatic opportunities for more natural, accurate and expressive communication. Our final goal is to create animated conversational agents that produce appropriate language-specific conversational behaviors, including accurate visible speech, emotional expressions, facial gestures (head nods, eye contact) and combined facial and upper body gestures. Possible applications of natural and expressive Talking Agents span a continuum from human-machine communication services to learning/teaching tutors, animated talking toys, services for the disabled people, and so on. To date, Baldi has achieved success as a language tutor for deaf children [2], children with autism [3], and adults learning English as a second language [24]. We look forward to similar successes for Baldini.

4. REFERENCES

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

[2] Massaro D.W., Cohen M.M., Beskow, J., Cole, R.A., Developing and evaluating conversational agents. In Cassell J., Sullivan J., Prevost S., Churchill E. (Eds.) *Embodied conversational agents*. MIT Press, 2000.

[3] Bosseler, A., Massaro, D.W. *Development and Evaluation of a Computer-Animated Tutor for Vocabulary and Language Learning in Children with Autism*. Submitted to Journal of Autism Research, 2002.

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

[5] 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.

[6] 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.

[7] 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.

[8] Pelachaud C., Magno-Caldognetto E., Zmarich C., Cosi P., "An approach to an Italian Talking Head", In *Proceedings of Eurospeech 2001*, Aalborg, Denmark, Sep 3-7 2001, pp. 1035-1038.

[9] Pelachaud C., Magno-Caldognetto E., Zmarich C., Cosi P., "Modelling an Italian Talking Head". In *Proceedings of AVSP-2001, Audio Visual Speech Processing Workshop*, Aalborg, Denmark, Sep 7-9 2001, pp. 72-77.

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

[14] FESTVOX: Alan W Black (awb@cs.cmu.edu), Kevin A. Lenzo (lenzo@cs.cmu.edu), Speech Group at Carnegie Mellon University. Website: http://www.festvox.org/.

[15] Cosi P., Avesani C., Tesser F., Gretter R., Pianesi F., "On the Use of Cart-Tree for Prosodic Predictions in the Italian Festival TtS". 2002, (in press).

[16] SCHEME, Computer Programming Language. Website: http://www-swiss.ai.mit.edu/~jaffer/Scheme.html.

[17] Dutoit T., Leich H., "MBR-PSOLA: Text-To-Speech Synthesis based on an MBE Re-Synthesis of the Segments Database". In *Speech Communication*, Elsevier Publisher, November 1993, vol. 13, n°3-4.

[18] Cosi P., Tesser F., Gretter R.. Avesani C., "Festival Speaks Italian!", In *Proceedings of Eurospeech 2001*, Aalborg, Denmark, Sep 3-7 2001, pp. 509-512.

[19] Fourcin A.J., Harland G., Barry W., Hazan V., Editors, *Speech Input and Output Assessment, Multilingual Methods and Standards*, Ellis Horwood Books in Information Technology, 1989.

[20] CSLU Labelling Guide, website: http://cslu.cse.ogi.edu/corpora/corpPublications.html.

[21] Cosi P., Hosom J.P., "High Performance 'General Purpose' Phonetic Recognition for Italian". *In Proceedings ICSLP-2000*, *International Conference on Spoken Language Processing*, Beijing, Cina, 16-20 October, 2000, Vol. II, pp. 527-530.

[22] Cosi P., Hosom J.P., Valente A., "High Performance Telephone Bandwidth Speaker Independent Continuous Digit Recognition", In *Proceedings Automatic Speech Recognition and Understanding Workshop*, Madonna di Campiglio, Trento (Italy), December 9-13, 2001.

[23] 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*, *International Conference on Spoken Language Processing*. Sydney, Nov 30-Dec 4, 1998, Vol. 7, pp. 3221-3224.

[24] Light, J., Massaro, D.W. Learning to perceive and produce non-native speech. Unpublished paper, 2001.