ITALIAN CHILDREN'S SPEECH RECOGNITION WITH APPLICATION TO INTERACTIVE BOOKS AND TUTORS

Piero Cosi*, Bryan Pellom**
* Istituto di Scienze e Tecnologie della Cognizione - Sezione di Padova "Fonetica e Dialettologia"
Consiglio Nazionale delle Ricerche , Via G. Anghinoni, 10 - 35121 Padova, Italy
**Center for Spoken Language Research
University of Colorado at Boulder - Boulder, Colorado, USA
cosi@pd.istc.cnr.it, pellom@cslr.Colorado.edu

ABSTRACT

This work represents a joint collaboration between the Center for Spoken Language Research (CSLR) at the University of Colorado and the Institute of Cognitive Sciences and Technologies of the National Research Council located in Padova Italy. This work was conducted with the specific goals of developing improved recognition of children's speech in Italian and the installation and integration of the children's speech recognition models into the Italian Literacy Tutor system. Specifically, children's speech recognition research for Italian was conducted using the ITC-irst Children's Speech Corpus (Giuliani & Gerosa, 2003). Using the University of Colorado SONIC large vocabulary speech recognition system, we demonstrate a phonetic recognition error rate of 13.5% for a system which incorporates Vocal Tract Length Normalization (VTLN), Cepstral variance normalization, Speaker-Adaptive Trained phonetic models, as well as iterative unsupervised Structural MAP Linear Regression (SMAPLR). These new acoustic models have been incorporated within an Italian version of the Colorado Literacy Tutor system.

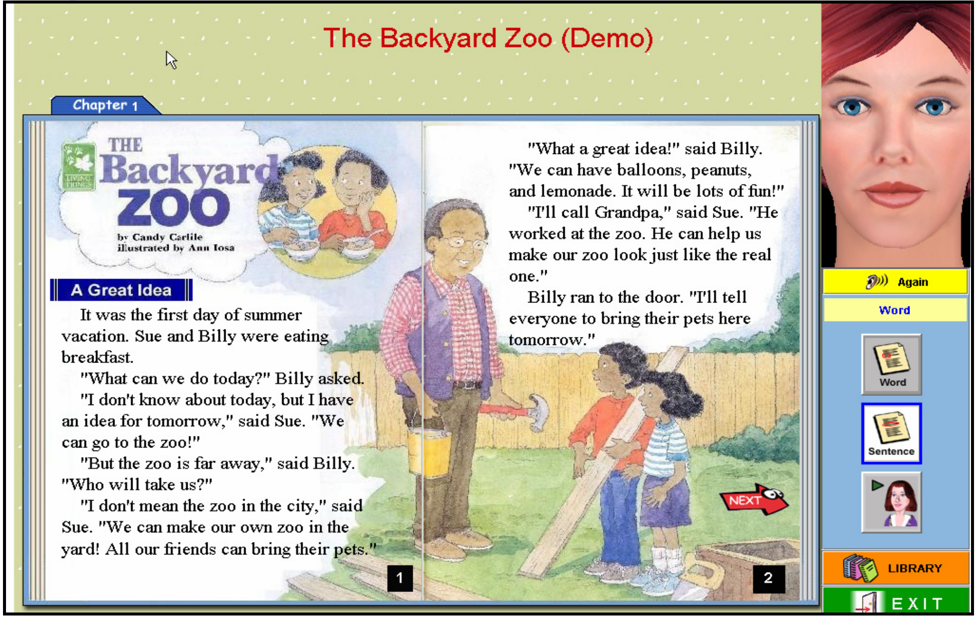
1. THE COLORADO LITERACY TUTOR

The Colorado Literacy Tutor (CLT) is a technology-based literacy program, designed on the basis of cognitive theory and scientifically motivated reading research, which aims to improve literacy and student achievement in public schools. The goal of the Colorado Literacy Tutor is to provide computer-based learning tools that will improve student achievement in any subject area by helping students learn to read fluently, to acquire new knowledge through deep understanding of what they read, to make connections to other knowledge and experiences, and to express their ideas concisely and creatively through writing. A second goal is to scale up the program to both state and national levels in the U.S. by providing accessible, inexpensive and effective computer-based learning tools.

The CLT project consists of four tightly integrated components: Managed Learning Environment, Foundational Reading Skills Tutors, Interactive Books, and Latent-Semantic Analysis (LSA)-based comprehension training (Steinhart 2001; Deerwester et al., 1990; Landauer and Dumais, 1997). A key feature of the project is the use of leading edge human communication technologies in learning tasks. The project has become a test bed for research and development of perceptive animated agents that integrate auditory and visual behaviors during face-to-face conversational interaction with human learners. The project enables us to evaluate component technologies with real users—students in classrooms—and to evaluate how the technology integration affects learning using standardized assessment tools.

Within the CLT, Interactive Books are the main platform for research and development of natural language technologies and perceptive animated agents. Figure 1 shows a page of an Interactive Book. Interactive Books incorporate speech recognition, spoken dialogue, natural language processing, and computer animation technologies to enable natural face-to-face conversational interaction with users. The integration of these technologies is performed using a client-server architecture that provides a platform-independent user interface for Web-based delivery of multimedia learning tools. Interactive Book authoring tools are designed for easy use by project staff, teachers and students to enable authors to design and format books by combining text, images, videos and animated characters. Once text and illustrations have been imported or input into the authoring environment, authors can orchestrate interactions between users, animated characters and media objects. Developers can populate illustrations (digital images) with animated characters, and cause them to converse with each other, with the user, or speak their parts in the stories using naturally recorded or synthetic speech. A mark up language enables authors to control characters' facial expressions and gestures while speaking. The authoring tools also enable authors to pre-record sentences and/or individual words in the text as well as utterances to be produced by animated characters during conversations.

Interactive Books enable a wide range of user and system behaviors. These include having the story narrated by animated characters, having conversations with animated characters in structured or mixed-initiative dialogues, having the student read out loud while words are highlighted, enabling the student to click on words to have them spoken by the agent or to have the agent interact with the student to sound out the word, having the student respond to questions posed by the agent either by clicking on objects in images or saying or typing responses, and having the student produce typed or spoken story summaries which can be analyzed for content using natural language processing techniques.



It was the first day of summer vacation. Sue and Billy were eating breakfast.

"What can we do today?" Billy asked.

"I don't know about today, but I have an idea for tomorrow," said Sue. "We can go to the zoo!"

"But the zoo is far away," said Billy. "Who will take us?"

"I don't mean the zoo in the city," said Sue. "We can make our own zoo in the yard! All our friends can bring their pets."

"What a great idea!" said Billy. "We can have balloons, peanuts, and lemonade. It will be lots of fun!"

"I'll call Grandpa," said Sue. "He worked at the zoo. He can help us make our zoo look just like the real one."

Billy ran to the door. "I'll tell everyone to bring their pets here tomorrow."

Figure 1: An example interactive book

2. BASELINE SPEECH RECOGNITION AND READING TRACKING ALGORITHM

The CLT uses the SONIC speech recognition system as a basis for providing real-time recognition of children's speech (Pellom, 2001; Pellom and Hacioglu, 2003; Hagen et al. 2004)[1] . The recognizer implements an efficient time-synchronous, beam-pruned Viterbi token-passing search through a static re-entrant lexical prefix tree while utilizing continuous density mixture Gaussian Hidden Markov Models (HMMs). The recognizer uses PMVDR cepstral coefficients (Yapanel and Hansen, 2003) as its feature representation. Children's acoustic models were estimated from 46 hours of audio from the CU Read and Prompted Children's Speech Corpus (Hagen et al., 2003) and the OGI Kids' speech corpus (Shobaki et al., 2000).

During oral reading, the speech recognizer models the story text using statistical n-gram language models. This approach gives the recognizer flexibility to insert/delete/substitute words based on acoustics and to provide accurate confidence information from the word-lattice. The recognizer receives packets of audio and automatically detects voice activity. When the child speaks, the partial hypotheses are sent to a reading tracking module. The reading tracking module determines the current reading location by aligning each partial hypothesis with the story text using a Dynamic Programming search. In order to allow for skipping of words or even skipping to a different place within the text, the search finds words that when strung together minimize a weighted cost function of adjacent word-proximity and distance from the reader's last active reading location. The Dynamic Programming search additionally incorporates constraints to account for boundary effects at the ends of each partial phrase.

Hagen et al. (2004a) describes more recent advances made to both acoustic and language modeling for oral-reading recognition of children's speech. Specifically, that work describes the use of cross-utterance word history modeling, position-sensitive dynamic n-gram language modeling, as well as vocal tract length normalization, speaker-adaptive training, and iterative unsupervised speaker adaptation for improved recognition. The final system was shown to have an overall word error rate of 8.0%. More recently the errors made by this baseline system have been investigated and work has been completed to provide for detection oral reading miscues (Lee et al., 2004b).

3. TRANSITIONING THE COLORADO LITERACY TUTOR TO ITALIAN

More recently Cosi (2004) describes initial research towards development of an Italian version of the Literacy Tutor described in Section 2.0. Under this proposed ISTC Reading Project, research will be conducted within the Italian language to provide improved student achievement in schools through the development of educational tools and technologies that assist students in learning to read and comprehend text. This work builds on tools developed by CNR such as the Italian FESTIVAL version (Cosi etal., 2001) and the LUCIA Mpeg-4 facial animation system (Cosi etal., 2003). The research therefore is focused around new state-of-the-art technologies for facial animation, speech synthesis, speech recognition, and text comprehension which are geared to support students within Italy.

[1] SONIC is freely downloadable for research use from (http://cslr.colorado.edu)

4. ITALIAN CHILDREN'S SPEECH RECOGNITION

4.1 Training Data and Initial System Port to Italian

The English version of the Colorado Literacy Tutor has been trained on speech data from over 1800 children aged 8-15 representing over 50+ hours of training audio. For Italian Children's speech recognition, we have used the ITC-irst Children's speech corpus which consists of 87 children aged 7 through 13 who were native speakers from the region in the north of Italy. Each child provided approximately 50-60 read sentences which were extracted from age-appropriate literature. The corpus was divided into a training set consisting of 67 speakers. The speaker labels for the training are provided in Table 1.

f0011	f0012	f0013	f0014	f0015
f0016	f0017	f0018	f0019	f0020
f0021	f0022	f0023	f0024	f0025
f0026	f0027	f0028	f0029	f0030
f0031	f0032	f0033	f0034	f0035
f0036	f0037	f0038	f0039	f0040
f0041	f0042	m0011	m0012	m0013
m0014	m0015	m0016	m0017	m0018
m0019	m0020	m0021	m0022	m0023
m0024	m0025	m0026	m0027	m0028
m0029	m0030	m0031	m0032	m0033
m0034	m0035	m0036	m0037	m0038
m0039	m0040	m0041	m0042	m0043
m0044	m0045			

Table 1: Training set speakers from the ITC-irst Children's Speech Corpus.

The SONIC speech recognition system was ported from U.S. English (adult 16 kHz microphone speech) to Italian kids models in the following manner. First, a phone mapping between target phonemes in Italian and U.S. English phonemes was determined. This mapping is used to provide initial Viterbi alignments on the training data. The Viterbi alignments are used to boot-strap the acoustic models into Italian. The phonetic mapping from Italian to U.S. English (Sphinx-II) phonemes is shown in Table 2.

It	U.S.	It	U.S.	It	U.S.	It	U.S.
a	AA	e	EY	i	IY	m	M
a1	AA	E	EH	i1	IY	n	N
b	B	e1	EY	j	Y	nf	NG
d	D	E1	EH	J	N	ng	NG
dz	ZH	f	F	k	K	o	OW
dZ	JH	g	G	l	L	O	AW
o1	OW	O1	AW	p	P	r	R
s	S	S	SH	t	T	ts	TS
tS	CH	u	UW	u1	UW	v	V
w	W	z	Z				

Table 2: Mapping of phonemes from Italian to U.S. English for system bootstrapping.

Given the phonetic mapping, an initial orthographic transcription in Italian and an Italian pronunciation dictionary the system first determines an initial Viterbi alignment of the acoustic training data. The Viterbi alignments provide the recognizer with an association to frames to states within the Hidden Markov Model (HMM). For this work, each phoneme is represented using a 3-state HMM model. Once the Viterbi alignment is determined, decision-tree state-clustered triphone HMM models are estimated using the SONIC system. Decision-tree splitting questions were designed specifically for Italian. Between 6-24 Gaussian mixtures per state are determined based on the amount of available training data. Given the initial acoustic model trained from Italian children's speech, the Viterbi alignment and retraining process are repeated to sequentially provide improved data alignments as well as improved acoustic models.

We have utilized this boot-strapping approach for a number of languages at the Center for Spoken Language Research (CSLR). These have included French, German, Spanish, Portuguese, Russian, Arabic, Korean, Japanese and Turkish. In all cases we have found that the initial selection of the phoneme mappings (e.g., as shown in Table 2), does not greatly impact the final error rate of the resulting recognition system.

4.2 Italian Children's Speech Database

Phonetic recognition experiments were conducted using the remaining 20 speakers from the ITC-irst Children's Speech Corpus. The speaker labels used for testing are shown in Table 3. For phonetic recognition we utilized the phoneme set shown in Table 4 consisting of 40 primary units. Results for phonetic recognition are presented using this 40 phoneme set as well as a reduced 33 unit set which does not take into account errors made between accented and non-accented vowels (e.g., "a" with "a1" and "o" with "o1").

f0001	f0002	f0003	f0004	f0005
f0006	f0007	f0008	f0009	f0010
m0001	m0002	m0003	m0004	m0005
m0006	m0007	m0008	m0009	m0010

Table 3: Test set speakers from the ITC-irst Children's Speech Corpus.

SAMPA	Example	SAMPA	Example
i	p**i**ni	i1	cos**ì**
e	v**e**lo	e1	merc**é**
E	asp**e**tto	E1	caff**è**
a	v**a**i	a1	bont**à**
o	p**o**lso	o1	R**o**ma
O	c**o**sa	O1	per**ò**
u	p**u**nta	U1	pi**ù**
j	p**i**ume	w	q**u**ando
K	**c**aldo	g	**g**atto
p	**p**era	B	**b**otte
T	**t**orre	d	**d**ente
ts	pi**zz**a	dz	**z**ero
tS	pe**c**e	dZ	ma**g**ia
m	**m**ano	n	**n**ave
ng	i**n**gordo	nf	a**n**fora
J	le**gn**a	L	so**gl**ia
l	pa**l**o	r	**r**emo
f	**f**aro	v	**v**ia

s	**s**ole	z	pe**s**o
S	**sc**i	SIL	silence

Table 4: Phoneme Set used for Italian Children's Speech Recognition.

For each experiment, a 3-gram phonetic language model was estimated from the resulting phonetic sequences from the phonetically aligned training data. The training data consists of 4,162 utterances. The 3-gram language model was estimated using the CMU/Cambridge Statistical Language Modeling Toolkit.

4.3 Experiments

In each experiment we utilize the phonetic sequences obtained by Viterbi alignment of the orthographic transcription of the test data in terms of phonetic sequence alignment. The phonetic aligner within SONIC allows for automatic detection and insertion of silence symbols during natural speaker pause in addition to automatically selecting the best pronunciation for a word given a set of alternative pronunciations in the Italian lexicon. Ideally, one would prefer to have a hand-labeled corpus which has been corrected at the phonetic level to take into account natural insertions, deletions and substitutions of phonetic units.

4.3.1 Phonetic Recognition of Children's Speech with Adult Models

In our first experiment, we wish to understand the phonetic error rate of a mismatched system (e.g., one trained on adult speech used to recognize children's speech). For this experiment we trained adult Italian acoustic models using the ITC-IRST APASCI speech corpus. APASCI is an Italian speech database recorded in an insulated room with a Sennheiser MKH 416 T microphone. The database contains 5,290 phonetically rich sentences in addition to 10,800 isolated digits (more than 10 hours of speech). The speech material was read by 100 Italian speakers (50 male and 50 female).

In this experiment, we perform incremental unsupervised structural MAP linear regression on the acoustic models given the confidence tagged system output (Siohan 2001, 2002). The means/variances are adapted using SMAPLR after each decoding pass. Results are shown in Table 5.

Previous research has shown that frequency warping prior to feature extraction can assist in reducing the mismatch between children's speech and adult acoustic models. In the next experiment we combine both SMAPLR (adaptation in the model space) with VTLN (adaptation in the feature space) to further improve the recognition of children's speech when only adult acoustic models are available. Results are shown in Table 6.

System Description	PER (40 units)	PER (33 units)
Speaker-Independent Adult	46.9%	41.1%
+ SMAPLR Pass 1	36.9%	31.3%
+ SMAPLR Pass 2	33.8%	28.3%
+ SMAPLR Pass 3	32.1%	26.8%
+ SMAPLR Pass 4	31.2%	25.9%
+ SMAPLR Pass 5	30.5%	25.3%

Table 5: Phonetic Error Rate (PER) as a function of SMAPLR adaptation iteration

System Description	PER (40 units)	PER (33 units)
Speaker-Independent Adult	46.9%	41.1%
+ VTLN/SMAPLR Pass 1	32.9%	27.6%
+ VTLN/SMAPLR Pass 2	30.1%	24.9%
+ VTLN/SMAPLR Pass 3	28.9%	23.8%

+ VTLN/SMAPLR Pass 4	28.2%	23.2%
+ VTLN/SMAPLR Pass 5	27.8%	22.9%

Table 6: Phonetic Error Rate (PER) as a function of SMAPLR adaptation iteration with VTLN applied to the children's data during feature extraction.

4.3.2 Viterbi-based Training of Italian Children's Speech Models

A total of 4 Viterbi alignment and acoustic model retraining passes were made on the training data. We first wished to determine if the models have sufficiently converged on the final best alignments of the training data. In Table 7, we show phonetic error rate as a function of model-alignment pass. We can see that 4 alignment passes is sufficient to achieve system convergence. What is interesting is that the initial alignments provided by the U.S. English phonetic mapping are only 10% relative worse than the final alignments obtained from the final Italian children's acoustic models.

System Description	PER (40 units)	PER (33 units)
Alignment Pass 0	24.8%	20.2%
Alignment Pass 1	23.5%	18.9%
Alignment Pass 2	22.9%	18.3%
Alignment Pass 3	22.6%	18.1%

Table 7: Phonetic Error Rate (PER) as a function of Viterbi alignment pass on the training data. Results are shown for the baseline 40 phonetic unit system and the reduced 33 unit system. Alignment Pass #0 is obtained using U.S. English acoustic models while Alignment Pass 1-3 are obtained using Italian Children's models estimated from the previous Viterbi data-alignment.

4.3.3 Experiments with Baseline Italian Children's Acoustic Models

We conducted an initial set of experiments to estimate the phonetic error rate for the baseline Italian children's models. In this experiment, perform incremental unsupervised structural MAP linear regression on the acoustic models given the confidence tagged system output (Siohan 2001, 2002). During each algorithm iteration, the means and variances are adapted using the SMAPLR algorithm. Results of this experiment are shown in Table 8. Typically systems developed at CSLR converge to a minimum error rate after 3 iterations of the SMAPLR algorithm. For phonetic recognition, we see that several additional iterations are required and still it is unclear as to exactly how many iterations are needed to achieve the lowest possible phonetic recognition error rate. This will be the subject of a future investigation.

System Description	PER (40 units)	PER (33 units)
Speaker-Independent	22.6%	18.1%
+ SMAPLR Pass 1	20.4%	16.0%
+ SMAPLR Pass 2	19.7%	15.3%
+ SMAPLR Pass 3	19.2%	14.9%
+ SMAPLR Pass 4	19.1%	14.7%
+ SMAPLR Pass 5	18.8%	14.6%

Table 8: Phonetic Error Rate (PER) as a function of SMAPLR adaptation iteration

4.3.4 Vocal Tract Length Normalization (VTLN)

We have considered improving our baseline children's acoustic models by performing vocal tract normalization for each child in the training set and to also perform VTLN frequency warp factor estimation for each test speaker. Frequency warping is applied in the range of 0.88 to 1.12 based on a children's speech acoustic model consisting of 1 Gaussian per clustered HMM state. The VTLN function determines the warping factor which maximizes the likelihood of the test data. A single VTLN warp factor is estimated for each speaker during training and test. In addition to VTLN, this experiment also includes cepstral variance normalization. Here, the variances of the cepstral parameters are adjusted to have unity variance. The normalization term is estimated using all the available adaptation data from the speaker. We can see from Table 9, that incorporating VTLN reduces the phonetic error rate from 18.8% to 18.0 for the 40 phonetic unit system and from 14.6% to 13.9% for the reduced 33 phonetic unit system.

System Description	PER (40 units)	PER (33 units)
Speaker-Independent	22.6%	18.1%
+ VTLN/SMAPLR Pass 1	18.9%	14.6%
+ VTLN/SMAPLR Pass 2	18.4%	14.1%
+ VTLN/SMAPLR Pass 3	18.2%	13.9%
+ VTLN/SMAPLR Pass 4	18.0%	13.7%
+ VTLN/SMAPLR Pass 5	18.0%	13.7%

Table 9: Phonetic Error Rate (PER) as a function of SMAPLR adaptation iteration for acoustic models which have been normalized by vocal tract length and variance normalized cepstral parameters.

4.3.5 Speaker Adaptive Trained (SAT) Models

Speaker Adaptive Training (SAT) attempts to remove speaker-specific characteristics from each of the training speakers in order to build better speaker-independent acoustic models. With the SONIC speech recognition system, we implement SAT by estimating a linear feature-space transformation (one transform per training speaker). The transform is estimated to maximize the likelihood of the training data to the VTLN/variance normalized acoustic model described in 4.3.D. During testing, the VTLN warp factor and cepstral variance normalization factor is estimated along with a Constrained MLLR (feature-space) transform. Decoding is then performed using the SAT trained models. Results of this final system are shown in Table 10. We see that SAT further reduces the phonetic recognition error rate from 18.0% to 17.9% for the 40 unit system and from 13.7% to 13.5% for the reduced 33 unit system.

System Description	PER (40 units)	PER (33 units)
Speaker-Independent	22.6%	18.1%
+ VTLN/SAT/SMAPLR Pass 1	18.7%	14.4%
+ VTLN/SAT/SMAPLR Pass 2	18.2%	13.9%
+ VTLN/SAT/SMAPLR Pass 3	18.0%	13.7%
+ VTLN/SAT/SMAPLR Pass 4	17.9%	13.6%
+ VTLN/SAT/SMAPLR Pass 5	17.9%	13.5%

Table 10: Phonetic Error Rate (PER) as a function of SMAPLR adaptation iteration

5. DISCUSSION

We have described our initial port of the SONIC large vocabulary speech recognition system to Italian for use in children's speech recognition. Using the ITC-irst Children's

Speech Database, a phonetic recognition error rate of 20.9% was achieved for first-pass speaker-independent recognition using a phonetic inventory of 40 units. Using a collapsed representation of 33 units, an error rate of 18.1% was demonstrated.

By utilizing advanced strategies for speech recognition decoding including Vocal Tract Length Normalization (VTLN) applied to the children's data; Structural MAP Linear Regression (SMAPLR); and Speaker Adaptive Training, it was demonstrated that the phonetic recognition error rate could be reduced to 17.9% for the 40 unit system and 13.5% for the reduced 33 unit system.

While the error rate for the current children's system is the lowest reported on that corpus (compare to 22.7% for the work of (Giuliani & Gerosa, 2003), there still exists a significant performance gap for acoustic models which have been trained on adult speech but used to decode children's speech. We investigated several means to reduce such mismatches including applying VTLN to the children's data to minimize the mismatch to the adult trained models and also applied iterative SMAPLR adaptation to shift the means and variances of the adult models to better match the children's voice characters. With both VTLN (feature-space transform) and SMAPLR (model-space transform) are applied to the mismatched condition, we see that the final system as an error rate of 27.8% for the 40 phonetic unit system and 22.9% for the reduced 33 unit system.

6. CONCLUSIONS

We have successfully ported the SONIC large vocabulary speech recognition system from English to Italian and have begun to consider the problem of optimization of the speech recognition system for Italian children's speech. The acoustic models developed were successfully integrated into the Colorado Literacy Tutor software and used to enable reading tracking in Italian for children's interactive books. This work will provide initial foundations for continued research towards development of an Italian Literacy Tutor.

7. FUTURE WORK

In the short-term, we will consider model-based transformations for mapping between adult acoustic models and children's acoustic voice data. Prior work has mainly focused on applying feature-based transformations to children's speech to map onto adult speech characteristics. We wish to estimate transforms between adult and children's acoustic spaces in a language-independent manner so that children's acoustic models can be rapidly built from existing adult audio databases in several languages.

REFERENCES

P. Cosi, F. Tesser., R. Gretter, C. Avesani (2001), "Festival Speaks Italian!", *Proceedings Eurospeech 2001*, Aalborg, Denmark, September 3-7, 2001, pages 509-512.

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

P. Cosi, R. Delmonte, S. Biscetti, R. Cole, B. Pellom, S. van Vuuren (2004), "Italian Literacy Tutor: tools and technologies for individuals with cognitive disabilities", in *InSTIL/ICALL Symposium 2004*, Venice, Italy, June.

S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman. (1990), "Indexing by Latent Semantic Analysis", *Journal of the Society for Information Science,* vol. 41, no. 6, pp. 391-407.

D. Giuliani and M. Gerosa (2003), "Investigating Recognition of Children's Speech", *Proc. IEEE-ICASSP 2003*, Hong Kong.

A. Hagen, B. Pellom, and R. Cole (2003), "Children's Speech Recognition with Application to Interactive Books and Tutors", *Proc. ASRU-2003*, St. Thomas, USA.

A. Hagen, B. Pellom, S. Van Vuuren, R. Cole (2004a), "Advances in Children's Speech Recognition within an Interactive Literacy Tutor", *Proc. HLT-NAACL 2004*, Boston Massachusetts, USA.

K. Lee, A. Hagen, N. Romanyshyn, S. Martin, B. Pellom (2004b), "Analysis and Detection of Reading Miscues for Interactive Literacy Tutors", *Proc. 20th International Conference on Computational Linguistics (Coling)*, Geneva, Switzerland, August.

T. Landauer and S. Dumais (1997), "*A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge", Psych. Review*, Vol. 104, pp. 211-240.

B. Pellom (2001), *SONIC: The University of Colorado Continuous Speech Recognizer*, Technical Report TR-CSLR-2001-01, University of Colorado.

B. Pellom, K. Hacioglu (2003), "Recent Improvements in the CU SONIC ASR System for Noisy Speech: The SPINE Task", *Proc. IEEE-ICASSP 2003*, Hong Kong.

K. Shobaki, J.-P. Hosom, and R. Cole (2000), "The OGI Kids' Speech Corpus and Recognizers", *Proc. ICSLP-2000*, Beijing, China.

O. Siohan, C. Chesta, and C.-H. Lee (2001), "Joint Maximum a Posteriori Adaptation of Transformation and HMM Parameters", *Proc. IEEE Trans. on Speech & Audio Processing*, Vol. 9, No. 4, pp. 417-428.

O. Siohan, T. Myrvoll, and C.-H. Lee (2002) " Structural Maximum a Posteriori Linear Regression for Fast HMM Adaptation", Computer, Speech and Language, 16, pp. 5-24, January.