RECENTI SVILUPPI DI SONIC PER L'ITALIANO: RICONOSCIMENTO AUTOMATICO DEL PARLATO INFANTILE

Piero Cosi
ISTC-spfd CNR
Istituto di Scienze e Tecnologie della Cognizione - Sede di Padova "Fonetica e Dialettologia"
Consiglio Nazionale delle Ricerche - via Martiri della Libertà, 2 - 35127 Padova, Italia e
e-mail: piero.cosi@pd.istc.cnr.it

1. SOMMARIO

In questo lavoro vengono descritti i risultati dei più recenti esperimenti di riconoscimento automatico di parlato infantile effettuati, mediante l'utilizzazione del sistema denominato SONIC, su un corpus di parlato letto da bambini di età compresa fra i 7 e i 13 anni. Il corpus utilizzato è stato raccolto presso alcune scuole del Trentino da parte dell' *ITC-IRST* ora *FBK* (*Fondazione Bruno Kessler*), nell'ambito di un progetto europeo denominato *PF-STAR*. In particolare, completando alcuni esperimenti realizzati passato, si è voluto integrare i nuovi modelli di riconoscimento allenati su voci di bambini nella versione italiana del Colorado Literacy Tutor. Il tasso di errore di riconoscimento iniziale di 15.1% per un insieme di 33 unità fonetiche (21,8% considerando un insieme di 40 unità fonetiche) è stato successivamente ridotto al 12.2% (18,6% considerando 40 unità) utilizzando una combinazione delle più aggiornate tecniche di adattamento comprendenti la normalizzazione di lunghezza del tratto vocale (*Vocal Tract Lenght Normalization*, *VTLN*), la normalizzazione della varianza dei coefficienti Cepstrali (*Cepstral coefficients Variance Normalization*, *CVN*) e l'utilizzazione di modelli fonetici addestrati in modalità indipendente dal parlante utilizzando le più recenti strategie iterative denominate *Structural MAP Linear Regression* (*SMAPLR*) e *Speaker Adaprive Training* (*SAT*). Questo lavoro è la continuazione ed il completamento naturale di un precedente simile lavoro (Cosi & Pellom, 2005) condotto su un insieme limitato dello stesso corpus di dati.

2. INTRODUZIONE

Il *Colorado Literacy Tutor* (*CLT*)[1] (Cole et alii, 2003), un sistema tecnologicamente avanzato ed interattivo costituito da una serie di "*tool*" computerizzati per l'insegnamento/apprendimento della lingua inglese e progettato sulla base delle più recenti teorie cognitive, mira a migliorare il livello di apprendimento degli studenti delle scuole primarie. Semplificando notevolmente *CLT* consiste di quattro moduli fortemente integrati fra loro denominati *Managed Learning Environment*, *Foundational Reading Skills Tutors*, *Interactive Books*, e *Latent Semantic Analysis (LSA)* e una caratteristica fondamentale è data dall'inserimento e dall'utilizzazione, nei moduli realizzati per l'apprendimento, delle più recenti e innovative tecnologie della comunicazione.

In particolare i "libri interattivi" sono la piattaforma principale per la ricerca e lo sviluppo delle tecnologie sul linguaggio naturale e gli agenti animati. Incorporando, infatti, il riconoscimento automatico del parlato, il *Trattamento Automatico del Linguaggio* naturale

[1] *Colorado Literacy Tutor*: http://www.colit.org/

(*TAL*) e le più recenti e innovative tecnologie grafiche di animazione al computer mirano a rendere sempre più naturale l'esperienza dell'apprendimento mediante ausili tecnologici.

In Cosi et alii (Cosi et alii, 2004) vengono descritte le attività di ricerca iniziali rivolte allo sviluppo della versione italiana del *CLT*, l' "*Italian Literacy Tutor*" (*ILT*).

ILT sarà realizzato basandosi su alcuni tool sviluppati in questi anni all' *ISTC-CNR* quali: la versione italiana di *SONIC* per il riconoscimento automatico del parlato infantile (Pellom, 2001; Pellom & Hacioglu 2003; Hagen et alii, 2003; Hagen et alii, 2004)[2], la versione italiana di *FESTIVAL* (Cosi et alii, 2001) per la sintesi da testo scritto e, *LUCIA*, una faccia parlante *MPEG-4* (Cosi et alii, 2003) in grado di esprimersi emotivamente.

Parallelamente al *CLT* l' *ILT* sarà costituito da una serie di "tool" computerizzati per l'insegnamento e l'apprendimento dell'italiano come lingua madre (*L1*) o lingua seconda (*L2*). Questo progetto, risultato della collaborazione fra Università, Centri di Ricerca e Scuole Pubbliche, mira a migliorare il livello e la qualità dell'apprendimento scolastico degli studenti della scuole di primo livello, mediante l'utilizzo di un software educativo sviluppato per aiutare gli allievi ad imparare a leggere e a comprendere correttamente un testo scritto.

Questi tool di apprendimento hanno un'enorme potenzialità e possono essere utilizzati per:

- insegnare a leggere e a capire un testo, all'interno di un completo programma di lettura, cercando possibilmente di identificare in età precoce eventuali soggetti disabili;
- migliorare la qualità del processo di apprendimento degli allievi aiutandoli ad acquisire specifiche conoscenze ed abilità mediante una più efficace capacità di comprensione del testo e mediante nuove ed efficaci strategie di scrittura;
- insegnare una seconda lingua

Una caratteristica fondamentale dell' "*Italian Literacy Tutor*" è quindi lo sviluppo di specifici strumenti per l'insegnamento della lettura agli allievi con particolari carenze. Molti bambini hanno infatti problemi di lettura che, qualora non vengano precocemente risolti, causano a lungo termine notevoli conseguenze negative. I soggetti che non riescono a leggere fluentemente generalmente sono impiegati in lavori secondari, sono caratterizzati da una notevole "sottostima", e sono incapaci di raggiungere i risultati che la loro potenziale capacità intellettuale e creativa consentirebbe. Se i problemi di lettura sono diagnosticati in età pre-scolare o nei primi anni di scuola possono essere sicuramente superati e risolti. Sebbene sia ormai assodato che un programma specifico e personalizzato di sostegno alla lettura possa fornire un'efficace soluzione le scuole non hanno risorse sufficienti a soddisfare tutte le possibili richieste e necessità.

L' *Italian Literacy Tutor* è progettato per fornire una soluzione efficace a questo problema fornendo una serie di tool di apprendimento per migliorare le capacità di lettura degli allievi e per identificarne eventuali carenze.

---

[2] Il sistema di riconoscimento *SONIC* è liberamente disponibile per scopi di ricerca per merito dell' *Università del Colorado* (http://cslr.colorado.edu)

ILT integra due tipologie di strumenti per l'apprendimento, uno basato sulle tecnologie dell'animazione e del parlato, e l'altro basato sulle tecnologia della comprensione del linguaggio.

Il primo insieme di strumenti include i Libri Interattivi (*Interactive Books*) e i Tutors Lettori che sono concepiti per lavorare assieme all'interno di un programma comprensivo di lettura. I Libri Interattivi aiuteranno agli studenti ad imparare a riconoscere le parole, leggere velocemente e comprendere ciò che leggono. Essi forniscono un ambiente per apprendere che va da lettori principianti (che possono farsi raccontare le storie dai personaggi animati, e poi essere coinvolti in dialoghi con i personaggi per valutare e esercitare la comprensione), a lettori avanzati che sono in grado di leggere le storie e quindi di ricevere addestramento alla lettura. I Libri Interattivi serviranno ad individuare le abilità di lettura mancanti o deboli, e indicheranno i Tutors Lettori individualizzati che valuteranno e insegneranno queste abilità. Sono in fase di sviluppo una serie di esercizi progettati per insegnare le abilità di base (*Foundational Skills*) fondamentali per una corretta lettura del testo. Mediante questi esercizi gli allievi interagiscono con agenti animati per l'apprendimento delle nozioni di base di una determinata lingua (*L1*, *L2*) quali ad esempio la conoscenza dell'alfabeto, la discriminazione e la produzione dei suoni linguistici, la consapevolezza fonologica, il "suono" e le parole, la struttura sillabica (Figura 2). L'apprendimento di queste capacità si è dimostrato particolarmente efficace nell'insegnamento della lettura in soggetti dotati di particolari problemi.

3. SISTEMA DI RICONOSCIMENTO AUTOMATICO PER IL PARLATO INFANTILE

Come già detto precedentemente *ILT* utilizza il sistema di riconoscimento del parlato continuo denominato *SONIC*, sviluppato all'università del Colorado, come architettura di base per il riconoscimento in tempo reale del parlato infantile (Pellom, 2001; Pellom & Hacioglu 2003; Hagen et alii, 2003; Hagen et alii, 2004). Il riconoscitore implementa un'efficace strategia di ricerca (*time-synchronous, beam-pruned Viterbi token-passing*) mediante un albero rientrante statico di prefissi lessicali e utilizza modelli markoviani nascosti con misture di gaussiane a densità di probabilità continua, anche fra le parole (HMMs). A livello di *front-end* acustico il riconoscitore utilizza come vettore di informazione i coefficienti cepstrali *PMDVR* (Yapanel & Hansen, 2003) o quelli classici *MFCC*.

3.1. Dati di addestramento e" porting" iniziale per l'italiano

La versione per l'inglese americano di *SONIC*, utilizzata nel *CLT*, è stata allenata utilizzando il parlato di più di 1800 bambini fra gli 8 e I 15 anni per un totale di più di 50 ore di parlato per l'addestramento del sistema (Hagen et alii, 2003; Shobaki et alii, 2000). In un task di lettura ad alta voce, si è ottenuto un errore di riconoscimento rispettivamente dell' 8% e dell' 11,5% a seconda dell'implementazione off-line e real-time del sistema (Hagen et alii, 2004).

Per l'apprendimento della versione italiana del riconoscitore di parlato infantile è stata utilizzata la versione completa del corpus *ChildIt* realizzato dall' *ITC-irst* (ora "*Fondazione Bruno Kessler*" - *FBK*) (Gerosa et alii, 2007) che è stato realizzato con le registrazioni di

171 bambini (85 femmine e 86 maschi) di età compresa fra i 7 e i 13 anni, nativi del Trentino.

Per ogni bambino sono state registrate approssimativamente circa 50-60 semplici frasi lette da alcuni libri adeguati alla loro età. Seguendo il lavoro di Gerosa et alii, (Gerosa et alii, 2007), il corpus è stato diviso in un insieme di training di 129 bambini (64 femmine e 65 maschi) e un insieme di test di 42 bambini (21 femmine e 21 maschi) bilanciati per sesso ed età fra i 7 e i 13 anni. Le frasi di training e di test contenenti parole mal pronunciate o forti rumori sovrapposti sono state preventivamente escluse negli esperimenti che verranno descritti di seguito, mentre tutte le altre frasi, anche quelle annotate con fenomeni extra linguistici tipo rumori dovuti ai parlanti (respiri, risate o colpi di tosse, …), rumori generici non sovrapposti con il segnale vocale (rumore generico, parlato estraneo non trascritto) e suoni non verbali o pause piene, sono state incluse e solo la loro trascrizione fonetica derivata dalla corrispondente trascrizione ortografica è stata utilizzata in fase di training e test.

Il sistema di riconoscimento *SONIC* dell' *Università del Colorado* allenato per voci di adulti americani (16 kHz, parlato microfonico) è stato trasformato nella versione per voci infantili italiane nel modo seguente: in un prima fase si è determinato una mappatura fonetica fra i fonemi target italiani, considerando 40 unità, e quelli inglesi americani; successivamente, questa mappatura fonetica, è stata utilizzata per fornire un primo allineamento forzato ottenuto mediante algoritmo di *Viterbi* con i moduli di riconoscimento per l'inglese americano e questo allineamento è servito come *boot-strap* per il training vero e proprio dei modelli acustici per l'italiano. I fonemi target italiani e la loro mappatura con quelli inglesi americani è illustrata in Tabella 1.

Mediante la mappatura fonetica, la trascrizione ortografica delle frasi target e un lessico di pronuncia, il sistema effettua inizialmente l'allineamento forzato mediante algoritmo di *Viterbi* delle frasi di training fornendo al riconoscitore l'associazione fra i *frames* acustici e gli stati dei modelli di *Markov* nascosti (*HMM*) associati alle parole delle frasi di training. Negli esperimenti che seguono, ogni fonema è rappresentato da un modello *HMM* a tre stati. Una volta determinato l'allineamento, vengono stimati i modelli *HMM* sulla base di alberi di decisione binari (*decision-tree state-clustered triphone HMM*). In *SONIC*, le domande che vengono poste nell'albero di decisione binario possono essere formulate in modo automatico per massimizzare la verosimiglianza (*likelihood*) dell'insieme di dati di training e non sono quindi necessarie domande basate su un'approfondita conoscenza linguistica per il *porting* di una lingua su di un'altra.

Ad ogni stato sono state assegnate da 6 a 24 misture di Gaussiane a seconda del materiale di training disponibile e, una volta allenati i modelli acustici iniziali, si è proceduto poi sequenzialmente ad un successivo allineamento forzato di *Viterbi* ed ad un nuovo addestramento per migliorare via-via i modelli acustici finali. Nei paragrafi seguenti sono descritti una serie di esperimenti che ben illustrano le problematiche incontrate nello sviluppo di un sistema di riconoscimento di parlato infantile.

IT	US	Example
i	IY	p**i**ni
E	EH	asp**e**tto
o	OW	p**o**lso
u	UW	p**u**nta
k	K	**c**aldo
t	T	**t**orre
tS	TS	pe**c**e
ng	NG	a**n**gora
l	L	pa**l**o
s	S	**s**ole
e	EY	v**e**lo
a	AA	v**a**i
O	AW	c**o**sa
j	Y	p**i**ume
p	P	**p**era
ts	TS	pi**zz**a
m	M	**m**ano
J	N	le**gn**a
f	F	**f**aro
S	SH	**Sc**i

IT	US	Example
i1	IY	cos**ì**
E1	EH	caff**è**
o1	OW	R**o**ma
u1	UW	pi**ù**
g	G	**g**atto
d	D	**d**ente
dZ	JH	ma**g**ia
nf	NG	a**n**fora
r	R	**r**emo
z	Z	pe**s**o
e1	EY	merc**é**
a1	AA	bont**à**
O1	AW	per**ò**
w	W	q**u**ale
b	B	**b**otte
dz	ZH	**z**ero
n	N	**n**ave
L	L	So**gl**ia
v	V	**v**ia
SIL	SIL	silence

Tabella 1: insieme di fonemi (*SAMPA*) utilizzati per il riconoscimento di parlato infantile italiano e corrispondente mappatura sull'inglese americano per il *bootstrapping* del sistema.

3.2. Corpus ChildIt

Gli esperimenti di riconoscimento fonetico sono stati eseguiti sul corpus *ChildIt* facendo uso di 42 parlanti per il test del sistema ovviamente non inclusi in quelli utilizzati per l'addestramento dei moduli acustici. Per il riconoscimento fonetico l'insieme di fonemi utilizzato è consistito di 40 unità acustiche primarie (AUs) (vedi la Tabella 1). I risultati per il riconoscimento fonetico sono presentati facendo uso di questo insieme di 40 unità come pure di un insieme ridotto di 33 unità acustiche che non considera gli eventuali errori di riconoscimento riscontrati sulle vocali accentate o atone (per esempio, "a" con "a1", oppure "o" con "o1"), errori che non pregiudicherebbero la *performance* del sistema qualora si utilizzassero le parole, quali unità di riconoscimento, invece delle unità acustiche assieme ad uno specifico modello del linguaggio.

3.3. Esperimenti

In ogni esperimento sono state utilizzate le sequenze fonetiche ottenute tramite l'allineamento di *Viterbi* della trascrizione ortografica dei dati di test come trascrizione fonetica di riferimento. Il modulo per l'allineamento fonetico forzato realizzato all'interno di *SONIC* tiene in considerazione, oltre a selezionare la migliore pronuncia per una parola dato un insieme di pronunce alternative estratte da un dizionario lessicale italiano, anche del rilevamento e inserzione automatici del simbolo della pausa o silenzio. Ovviamente, nella migliore delle ipotesi, sarebbe preferibile disporre di un corpus etichettato e segmentato manualmente a livello fonetico per considerare eventuali inserzioni, cancellazioni e sostituzioni di unità fonetiche nella realizzazione effettiva delle frasi target

sia per il training che per il test. Per ognuno degli esperimenti descritti nei paragrafi seguenti, è stato inoltre stimato un modello fonetico del linguaggio a tri-grammi (Clarkson & Rosenfeld, 1997) a partire dalle sequenze fonetiche risultanti dai dati allineati di addestramento che consistono di 13765 espressioni.

3.3.1. Riconoscimento di parlato infantile con modelli acustici di parlato adulto

Nella nostra prima serie di esperimenti, si desiderava capire il tasso di errore fonetico di riconoscimento di un sistema mal adattato, cioè, un sistema addestrato su voci di parlanti adulti per riconoscere il parlato infantile. Si desiderava inoltre quantificare la riduzione degli errori che poteva essere ottenuta qualora si fossero utilizzati alcuni metodi di normalizzazione e di adattamento. Per questo esperimento sono stati utilizzati i modelli acustici per l'italiano addestrati su parlanti adulti mediante il corpus *APASCI* realizzato da *FBK* (ex *ITC-irst*). *APASCI* è un corpus di parlato italiano adulto registrato in una camera silente mediante microfono Sennheiser MKH 416 T. Il corpus contiene 5.290 frasi foneticamente ricche oltre a 10.800 cifre isolate (più di 10 ore di parlato). Il materiale vocale è stato letto da 100 parlanti (50 maschi e 50 femmine) italiani. E' stata utilizzata la procedura di *porting* dall'inglese americano all'italiano descritta nel paragrafo 3.1 e sono stati stimati i modelli acustici indipendenti dal parlante e quelli dipendenti dal genere maschile o femminile. Per ridurre il disallineamento fra i modelli di parlato adulto ed i dati acustici dei bambini, è stata applicata la regressione lineare strutturata massima a posteriori (*Maximum-A-Posteriori, MAP*) non supervisionata (*SMAPLR*), utilizzando l'uscita fonetica del riconoscitore opportunamente pesata mediante una misura di affidabilità (Siohan, 2002). Le medie e le varianze delle gaussiane del sistema sono state adattate utilizzando la procedura *SMAPLR* dopo ogni passaggio di decodifica e sono state utilizzate per ottenere un risultato fonetico migliore dil riconoscimento. I risultati sono indicati in Tabella 2 (a).

Ricerche precedenti inoltre hanno indicato che la normalizzazione della lunghezza del tratto vocale (*Vocal Tract Lenght Normalization, VTLN*) mediante deformazione dell'asse delle frequenze prima dell'estrazione del vettore delle caratteristiche acustiche può sicuramente essere di aiuto per la riduzione del non allineamento fra il parlato dei bambini ed i modelli acustici per gli adulti. In *SONIC* è implementato il metodo di deformazione dell'asse delle frequenze descritto in Welling et alii (Welling et alii,1999). La funzione *VTLN* determina il fattore di deformazione, necessario per far corrispondere i dati anatomici medi di adulti comparati a quelli dei bambini, che varia fra 0,88 e 1,12 per ogni parlante in modo tale da massimizzare la verosimiglianza (likelihood) dei dati di test. I risultati degli esperimenti che combinano *SMAPLR* e *VTLN* sono riassunti in Tabella 2 (b). Dalla tabella 2 (a) possiamo vedere che il tasso di errore fonetico iniziale è 39,2% per un sistema che consiste di 40 unità acustiche (AU) (31,1% per 33 AU) quando i modelli acustici addestrati su voci di adulti sono stati utilizzati per riconoscere il parlato infantile. Come ci si poteva attendere, inoltre, i modelli acustici addestrati unicamente su parlanti adulti femmine, forniscono un piccolo miglioramento rispetto a quelli indipendenti dal parlante - riducendo il tasso di errore fonetico iniziale a 36,8% e a 28,7% rispettivamente per 40 o 33 AU. L'adattamento mediante *SMAPLR* riduce ulteriormente il tasso di errore fonetico a 28,1% e 20,7% (vedi la Tabella 2a) rispettivamente per 40 o 33 unità acustiche.

La combinazione di *VTLN* nello spazio delle caratteristiche acustiche con *SMAPLR* nello spazio del modello riduce ancora il tasso di errore a 26,7% e a 19,3% (Tabella 2 (b)).

Riassumendo, può essere raggiunta infine una riduzione dell'errore relativo di quasi 32% combinando l'adattamento dello spazio-acustico (*SMAPLR*) e l'adattamento dello spazio delle caratteristiche (*VTLN*) ai modelli acustici addestrati esclusivamente su voci femminili adulte.

SMAPLR Adaptation	(a) Speaker Ind.		(b) Adult Female	
	PER 40 AU	PER 33 AU	PER 40 AU	PER 33 AU
First-Pass	39.2%	31.1%	36.8%	28.7%
+Adapt Iter. 1	31.7%	24.1%	29.6%	22.0%
+Adapt Iter. 2	29.7%	22.2%	27.8%	20.3%
+Adapt Iter. 3	28.9%	21.5%	27.0%	19.6%
+Adapt Iter. 4	28.4%	21.0%	26.5%	19.1%
+Adapt Iter. 5	28.1%	20.7%	26.5%	18.8%

Tabella 2. Tasso di errore fonetico per il riconoscimento del parlato di bambini (*PER*) in funzione della ripetizione dell'addestramento *SMAPLR* per i modelli acustici addestrati su voci di parlanti adulti indipendenti dal parlante (a) e di parlanti femminili (b) con adattamento *VTLN* e *SMAPLR*.

Nella sezione seguente, viene descritto lo sviluppo dei modelli acustici addestrati unicamente sul parlato di bambini per illustrare il grado di non allineamento che tuttora esiste fra i modelli adulti adattati alle voci di bambini ed i modelli infantili veri e propri.

3.3.2. Addestramento di Viterbi dei modelli di parlato infantile italiano

Come citato nella sezione 3.1, il metodo di *porting* di *SONIC* dall'inglese all'italiano fa affidamento su di un iniziale mappatura fonetica basata su una conoscenza linguistica specifica della differenza fra i fonemi target e i fonemi della lingua di partenza. Fino ad oggi, *SONIC* è stato utilizzato per il *porting* su quasi 20 lingue e l'esperienza ha indicato che l'accuratezza della mappatura iniziale ha un impatto minimo sul tasso di errore finale dei modelli acustici risultanti. In questo lavoro sono stati effettuati un numero totale di 6 allineamenti di Viterbi e di re-training dei modelli acustici per ottenere i modelli finali di riconoscimento di voci di bambini. In Tabella 3, è illustrato il tasso di errore di riconoscimento fonetico in funzione della ripetizione dell'allineamento ed è chiaro che 6 passaggi di allineamento sono sufficienti per raggiungere la convergenza del sistema.

Vale la pena di notare che i modelli di riconoscimento base per i bambini rivelano una riduzione del 10% del tasso di errore del riconoscimento fonetico se confrontato ai migliori modelli adulti adattati al parlato dei bambini (vedi Tabella 2).

3.3.3. Esperimenti di riconoscimento con modelli acustici di parlato infantile

Come descritto nella sezione 3.3.1, in modo simile a quello eseguito per gli esperimenti con i modelli allenati su parlato adulto, i modelli base allenati direttamente sulle voci di bambini sono stati estesi mediante l'introduzione dell'adattamento iterativo SMAPLR. I risultati di questo esperimento sono indicati in Tabella 4 (a).

Viterbi Training Step	Children's Acoustic Models	
	PER (40 AU)	PER (33 AU)
Align / Train Pass 0	24.4%	17.4%
Align / Train Pass 1	22.8%	15.9%
Align / Train Pass 2	22.1%	15.4%
Align / Train Pass 3	21.7%	15.1%
Align / Train Pass 4	21.7%	15.1%
Align / Train Pass 5	21.7%	15.0%
Align / Train Pass 6	21.8%	15.1%

Tabella 3: Tasso di errore sul riconoscimento fonetico (*PER*) in funzione dell'iterazione dell'allineamento di Viterbi per l'addestramento del modulo acustico di riconoscimento di parlato infantile sul corpus di parlato letto ChildIt dell' *ITC-IRST* (ora *FBK*). Si noti che all'iterazione iniziale (0) l'allineamento è ottenuto utilizzando i modelli acustici inglesi americani mentre nelle iterazioni 1-6 è ottenuto con i modelli acustici addestrati direttamente sul parlato di bambini.

Children's Speech Phonetic Recognition	(a) SMAPLR Adaptation		(b) SMAPLR & VTLN	
	PER 40 AU	PER 33 AU	PER 40 AU	PER 33 AU
First-Pass	21.8%	15.1%	21.8%	15.1%
+Adapt Iter. 1	20.3%	13.6%	19.0%	12.6%
+Adapt Iter. 2	19.9%	13.3%	18.7%	12.4%
+Adapt Iter. 3	19.8%	13.2%	18.7%	12.3%
+Adapt Iter. 4	19.8%	12.3%	18.7%	12.3%
+Adapt Iter. 5	19.8%	13.2%	18.7%	12.3%

Tabella 4: (a) tasso di errore di riconoscimento fonetico (PER) in funzione della ripetizione dell'adattamento *SMAPLR*; (b) tasso di errore di riconoscimento fonetico (PER) in funzione della ripetizione dell'adattamento SMAPLR a cui è stata aggiunta la normalizzazione del tratto vocale (*VTLN*).

Diversamente dagli esperimenti descritti in 3.3.1 con i modelli adulti, sono necessarie meno ripetizioni di adattamento per raggiungere il tasso di errore di riconoscimento fonetico più basso. L'adattamento acustico applicato ai modelli dei bambini riduce ulteriormente l'errore di quasi il 9%. Inoltre, come visto in precedenza, si voleva verificare se sarebbe stato possibile ottenere un miglioramento del sistema qualora le differenze del tratto vocale fossero rimosse fra i vari bambini nell'insieme di training e per far questo è stata applicata la normalizzazione del tratto vocale (*VTLN*) per ogni bambino appartenente all'insieme di training ed è stata applicata anche la normalizzazione *VTLN* previa stima del fattore di ditorsione dell'asse delle frequenze per ogni bambino appartenente all'insieme di test (Welling et alii, 1999). Possiamo vedere dalla Tabella 4 che incorporando la procedura *VTLN* il tasso di errore di riconoscimento fonetico si riduce dal 21,8% al 18,7% per il sistema con 40 unita acustiche e dal 15,1% al 12,3% per il sistema ridotto con 33 unità acustiche. Come già anticipato, i guadagni dovuti all'applicazione della procedura di *VTLN*

sono meno sostanziali in questa condizione di quanto non lo siano in condizioni di non allineamento sostanziale dei modelli (cioè, modello adulto con parlato infantile).

Da ultimo, la tecnica di adattamento denominata *Speaker Adaptive Training* (*SAT*) mira a rimuovere le caratteristiche specifiche di un parlante sui dati di training al fine di stimare i parametri dei modelli acustici indipendenti dai parlanti. In *SONIC* la procedura *SAT* viene realizzata mediante la stima di una singola trasformazione lineare nello spazio delle caratteristiche acustiche per ogni parlante dell'insieme di addestramento.

Questa funzione di trasformazione viene stimata con il vincolo di massimizzare la probabilità dei dati di addestramento una volta stimato il modello acustico dopo aver applicato la procedura di *VTLN*.

Durante la fase di test, il fattore di distorsione viene stimato mediante un unica funzione di trasformazione (*Maximum Likelihood Linear Regression*) nello spazio delle caratteristiche acustiche prima del riconoscimento. Questo sistema finale riduce il *PER* dal 21,8% al 18,6% per il sistema con 40 unità acustiche e dal 15,1% al 12,2% per il sistema ridotto con 33 unità, come illustrato in Tabella 5.

Italian Children's Speech Phonetic Recognition	**SMAPLR + VTLN + SAT**	
	PER 40 AU	**PER 33 AU**
First-Pass	**21.8%**	**15.1%**
+Adapt Iter. 1	**19.0%**	**12.5%**
+Adapt Iter. 2	**18.7%**	**12.3%**
+Adapt Iter. 3	**18.6%**	**12.2%**
+Adapt Iter. 4	**18.6%**	**12.2%**
+Adapt Iter. 5	**18.6%**	**12.2%**

Tabella 5: Errore di riconoscimento fonetico (*PER*) con modelli acustici di parlato infantile per un sistema in cui sono state applicate le procedure *SMAPLR*, *VTLN* e *Speaker Adaptive Training* (*SAT*) in funzione dell'iterazione di re-training.

3.3.4. Discussione degli esperimenti

Mentre il tasso di errore del sistema allenato su voci di bambini è paragonabile e addirittura migliore di quello ottenuto da sistemi simili sullo stesso corpus (ad esempio paragonabile al 22.7% ottenuto da un sistema analogo con 28 unità fonetiche come quello utilizzato in Giuliani & Gerosa, 2003), esiste ancora un significativo margine di miglioramento per un sistema che utilizzi modelli acustici allenati su parlato adulto e utilizzati per decodificare parlato infantile. Infatti quando sono state applicate entrambe le tecniche *VTLN* e *SMAPLR* in una condizione di disallineamento adulti/bambini il sistema finale ha ottenuto un tasso di errore fonetico del 19.3% dimostrando di ridurre l'errore fonetico iniziale del 28%. Ciò nonostante, persiste ancora un notevole 30% di differenza relativa fra l'utilizzazione di modelli acustici allenati su parlato adulto e modelli acustici allenati su parlato infantile per la decodifica di quest'ultimo.

Il tasso di errore di riconoscimento iniziale di 15.1% per un insieme di 33 unità fonetiche (21,8% considerando un insieme di 40 unità fonetiche) è stato successivamente ridotto al 12.2% (18,6% considerando 40 unità) utilizzando una combinazione delle più aggiornate tecniche di adattamento comprendenti la normalizzazione di lunghezza del tratto vocale (*VTLN*), la normalizzazione della varianza dei coefficienti Cepstrali (*Cepstral coefficients Variance Normalization, CVN*) e l'utilizzazione di modelli fonetici addestrati in modalità indipendente dal parlante utilizzando le più recenti strategie iterative denominate *Structural MAP Linear Regression* (*SMAPLR*) e *Speaker Adaprive Training* (*SAT*).

4. CONCLUSIONI E SVILUPPI FUTURI

Lo sviluppo di sistemi di riconoscimento di parlato infantile spesso si presenta come un compito di ardua soluzione a causa della spesso totale mancanza di risorse acustiche utilizzabili per l'allenamento dei modelli acustici. In questo lavoro, il sistema di riconoscimento denominato *SONIC* e sviluppato per l'inglese è stato adattato all'italiano ed in particolare è stato considerato il caso del parlato infantile di bambini compresi nella fascia di età compresa fra i 7 e i 13 anni.

Questi nuovi modelli acustici per il parlato infantile italiano sono stati incorporati nel *CLT* (*Colorado Literacy Tutor*), sviluppato al *CSLR* (*Centre for Speech and Language Research*) della *University of Colorado di Boulder*, per la lingua inglese, quale primo passo verso lo sviluppo della sua corrispondente versione italiana l'*Italian Literacy Tutor* (Cosi et alii, 2004), un sistema interattivo e personalizzato per l'apprendimento della lingua italiana. In particolare in Figura 1 è illustrata la videata iniziale di un libro interattivo utilizzato per insegnamento/apprendimento della lettura in italiano, in cui si può notare lin alto a destra l'assistente virtuale *LUCIA* (Cosi et alii, 2003) che può ad esempio leggere il testo ed in generale interagire con gli utenti durante il processo di apprendimento.

All'interno del *CLT e in futuro dell'ILT*, il riconoscimento del parlato viene utilizzato per seguire il livello raggiunto dal bambino nell'apprendimento della lettura, e può assisterlo nella rilevazione degli eventuali possibili errori, in generale allo scopo di fornire alcune informazioni utilizzabili in fase di misurazione della fluenza di lettura. Per migliorare il riconoscimento dei testi il sistema di riconoscimento costruisce un modello del linguaggio del testo del libro a tri-grammi al fine di fornire quella flessibilità necessaria per inserire/cancellare/sostituire le parole in funzione del modello acustico elaborato e per fornire inoltre delle misure di confidenza acustica calcolate dal lattice di ipotesi di parola fornito dal riconoscitore.

Quando il bambino parla, le ipotesi parziali vengono inviate al modulo di riconoscimento che determina la posizione attuale della lettura allineando ogni ipotesi parziale con la storia del testo utilizzando un algoritmo di ricerca basato sulla programmazione dinamica. Hagen et alii (Hagen et alii, 2004) descrive in dettaglio gli avanzamenti più recenti raggiunti sia nella modellizzazione acustica sia nella realizzazione di efficienti modelli del linguaggio per il riconoscimento della lettura di parlato infantile.

Di conseguenza, ILT utilizzerà in futuro un modulo per il controllo della storia/cronologia delle parole anche a cavallo di frase, nuovi modelli del linguaggio a trigrammi dinamici e adattabili alla posizione nel testo, come pure, a livello acustico, un modulo specifico per la

normalizzazione del tratto vocale, per l'addestramento adattativo ai parlanti, e per l'addestramento dei parlanti senza supervisione.

Il sistema risultante per l'inglese americano ha dimostrato di raggiungere un tasso di errore globale di riconoscimento di parola dell' 8,0%. In uno studio successivo, gli errori fatti da questo sistema di base sono stati analizzati ed utilizzati per migliorare lo sviluppo di un sistema in grado di riconoscere possibili errori di lettura (Lee et alii, 2004).

Basandosi su tutte queste ricerche, il sistema di riconoscimento SONIC è stato esteso per realizzare il tracciamento della lettura e l'analisi del segnale verbale utilizzando come unità acustiche delle unità di dimensioni più piccole della parola (Hagen & Pellom, 2005).

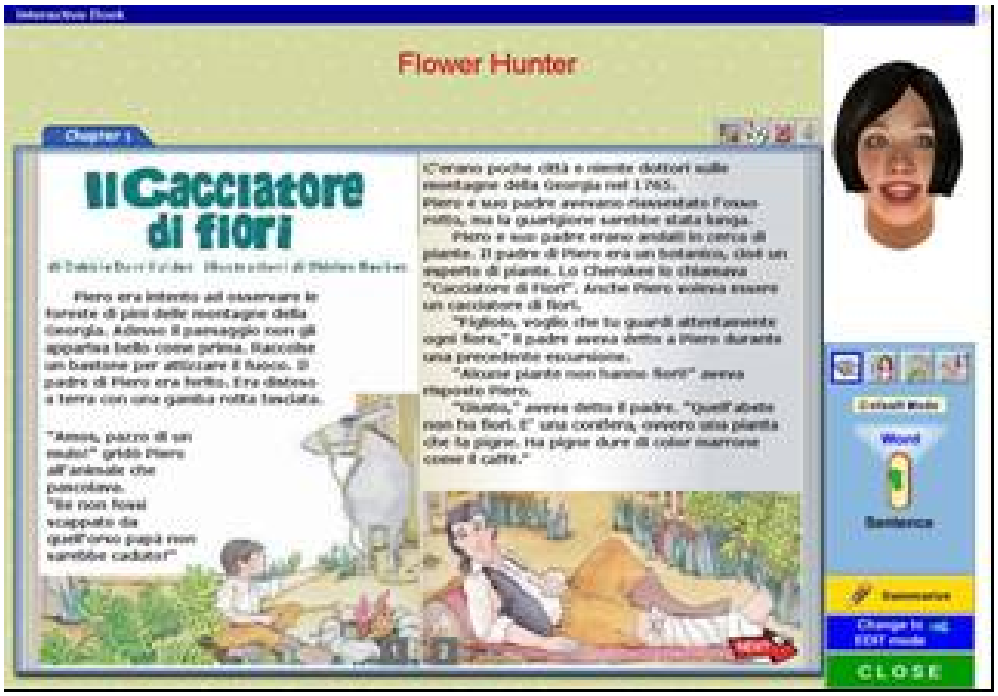

Figure 1: *Italian Literacy Tutor* (*ILT*): videata iniziale di un libro interattivo

5. RINGRAZIAMENTI

Un calorosissimo ringraziamento va all'intero gruppo di ricerca CSLR (*Computer Spoken Language Research*) dell'Università del Colorado e soprattutto a Bryan Pellom (ora in *Rosetta Stone*) per i suoi insostituibili suggerimenti e la sincera amicizia.

6. BIBLIOGRAFIA

Cole, R., van Vuuren S., Pellom B., et al. (2003), Perceptive Animated Interfaces: First Steps Toward a New Paradigm for Human Computer Interaction, in *Proceedings of the IEEE*, vol. 91, no. 9, Sept., 2003, 1391-1405.

Cosi, P., Tesser, F., Gretter, R. & Avesani C. (2001). Festival Speaks Italian!, in *Proceedings of Eurospeech 2001*, 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, in *Proceedings of Eurospeech 2003*, Geneva CH, 127-132.

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

Cosi, P. & Pellom, B. (2005), Italian Children's Speech Recognition For Advanced Interactive Literacy Tutors, *in CD-Rom Proceedings of INTERSPEECH 2005*, Lisbon, Portugal, 2005, 2201-2204.

Gerosa, M., Giuliani, D. & Brugnara, F. (2007), Acoustic Variability and automatic recognition of children's speech, *Speech Communication*, Vol. 49 (10-11), Oct. 2007, 847-860.

Hagen, A. & Pellom, B. (2005), A Multi-Layered Lexical-Tree Based Token Passing Architecture for Efficient Recognition of Subword Speech Units, *in 2nd Language & Technology Conference*, Poznan, Poland, April, 2005.

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

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

Lee, K., Hagen, A., Romanyshyn, N., Martin, S. & Pellom, B. (2004), Analysis and Detection of Reading Miscues for Interactive Literacy Tutors, in *Proceedings of 20th International Conference on Computational Linguistics (Coling)*, Geneva, CH, 2004.

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

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

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

Shobaki K., Hosom J.P., and Cole R. (2000), The OGI Kids' Speech Corpus and Recognizers, Proc. ICSLP, Beijing, China, 2000, Vol. IV, 564-567.

Yapanel, U.H. & Hansen, J.H.L. (2003), A New Perspective on Feature Extraction for Robust In-Vehicle Speech Recognition, *in Proceedings EUROSPEECH 2003*, Geneva, Switzerland, September 1-4, 2003, 1281-1284.

Welling, L., Kanthak, S. & Ney, H. (1999), Improved Methods for Vocal Tract Length Normalization in *Proceedings of ICASSP*, Phoenix, Arizona, USA, 1999.

Giuliani, D. & Gerosa, M. (2003), Investigating Recognition of Children's Speech, in *Proceedings of ICASSP*, Hong Kong, 2003.