# On the Development of Matched and Mismatched Italian Children's Speech Recognition Systems

*Piero Cosi*

Istituto di Scienze e Tecnologie della Cognizione, C.N.R.

piero.cosi@pd.istc.cnr.it

## Abstract

While at least read speech corpora are available for Italian children's speech research, there exist many languages which completely lack children's speech corpora. We propose that learning statistical mappings between the adult and child acoustic space using existing adult/children corpora may provide a future direction for generating children's models for such data deficient languages. In this work the recent advances in the development of the SONIC Italian children's speech recognition system will be described. This work, completing a previous one developed in the past, was conducted with the specific goals of integrating the newly trained children's speech recognition models into the Italian version of the Colorado Literacy Tutor platform. Specifically, children's speech recognition research for Italian was conducted using the complete training and test set of the FBK (ex ITC-irst) Italian Children's Speech Corpus (ChildIt). Using the University of Colorado SONIC LVSR system, we demonstrate a phonetic recognition error rate of 12,0% for a system which incorporates Vocal Tract Length Normalization (VTLN), Speaker-Adaptive Trained phonetic models, as well as unsupervised Structural MAP Linear Regression (SMAPLR).

**Index Terms**: children, ASR, Italian, adaptation.

## 1. Introduction

The Colorado Literacy Tutor (CLT) [1,2] is a technology-based literacy program, designed on the basis of cognitive theory and scientifically motivated reading research, which aims to improve student achievement in public schools. The CLT uses the University of Colorado SONIC speech recognition system as a basis for providing real-time recognition of children's speech [3-6][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, cross-word, mixture Gaussian Hidden Markov Models (HMMs). The recognizer uses PMVDR cepstral [7] or classical MFCC coefficients as its feature representation. To adapt the speech recognizer to better match the test condition SONIC implements several feature-based (CMS - cepstral mean subtraction, VTLN - vocal tract length normalization and CVN - cepstral variance normalization) and several direct (MAP - maximum a posteriori estimation) and indirect (ML - maximum likelihood) model based adaptation techniques.

Moreover, several supervised vs.unsupervised and block vs. Incremental modes of adaptation are possible. In the unsupervised case, the transcription is not known and should be estimated in some form; either as a single best string or a word lattice. In incremental adaptation the models are adapted as enough data becomes available, and the new models are used to decode the incoming data, which, in turn, is used to readapt the models. In block adaptation, the adaptation is started after all data is available. Within SONIC several adaptation schemes are considered:

- Maximum likelihood linear regression (MLLR):
  - incremental / block, (ii) single class / multiple class,
  - best string / word lattice
- Maximum a posterior linear regression (MAPLR):
  - block (ii) best string / word lattice
  - regression class tree.

By using SONIC, Hagen et al. [6] describes some advances made to both acoustic and language modeling for oral-reading recognition of children's speech using cross-utterance word history modeling, position-sensitive dynamic n-gram language modeling, as well as vocal tract length normalization, speaker-adaptive training, and unsupervised speaker adaptation for improved children's speech recognition. The resulting U.S. English system was shown to have an overall word error rate of 8.0%. In a later study, errors made by this baseline system were analyzed and used to inform the development of a system for detecting oral reading miscues [8]. Based on that work, the SONIC speech recognition system was extended to perform reading tracking and speech analysis using subword sized acoustic units [9].

## 2. Italian Children Speech ASR

### 2.1. Training Data and Initial System Port to Italian

The work presented here is the natural continuation and completion of a quite similar work [10] conducted on a limited set of the same speech data. The U.S. 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 [5,11]. For Italian Children's speech recognition, we have used the final release of the FBK ChildIt corpus [12] which consists of data collected from 171 children (85 females and 86 males) aged between 7 and 13 (from grade 2 up to grade 8) 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. Following the work in [13], the corpus was divided into a training set consisting of 129 speakers (64 females and 65 males) and a test set consisting of 42 speakers (21 females and 21 males) balanced by gender and aged between 7 and 13. Training and test sentences containing mispronunciation and noisy words were excluded in the following experiments while all other sentences, even those with annotated extra-linguistic phenomena like noises due to the speaker (lip smacks, breath, laugh, cough, …), generic noises non overlapping with speech (generic noise, untranscribed extraneous speech) and non

[1] The SONIC speech recognition system is available for research use from the University of Colorado (http://cslr.colorado.edu)

verbal sounds or filled pauses were included, and only the phonetic transcriptions of the prompt sentences were used for training and test.

The University of Colorado SONIC speech recognition system was ported from U.S. English (adult 16 kHz microphone speech) to Italian children's models in the following manner. First, a phonetic mapping between target phonemes in Italian and U.S. English phonemes was determined. Our primary phoneme set for Italian consists of 40 units. The phonetic 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 target phoneme set for Italian and phonetic mapping from Italian to U.S. English phonemes is shown in Table 1.

Given the phonetic mapping, an initial orthographic transcription 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. In SONIC, the decision-tree splitting questions can be formed in an automated fashion such that the likelihood of the training data is maximized. Thus hand-derived splitting questions based on expert linguistic knowledge are not needed for language porting. The resulting clustered states each were assigned 6 to 24 Gaussian mixtures 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.

Table 1. *Phoneme Set (SAMPA) used for Italian Children's Speech Recognition and mapping of phonemes from Italian (IT) to U.S. English (US) for system bootstrapping.*

| IT | US | Example |
|---|---|---|
| I | IY | p**i**ni |
| E | EH | aspetto |
| 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 | **co**sa |
| J | Y | pi**u**me |
| 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 | così |
| E1 | EH | caffè |
| o1 | OW | R**o**ma |
| u1 | UW | più |
| g | G | **g**atto |
| d | D | **d**ente |
| dZ | JH | ma**gi**a |
| 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 | **qu**ale |
| b | B | **b**otte |
| dz | ZH | **z**ero |
| n | N | **n**ave |
| L | L | Sog**li**a |
| v | V | **v**ia |
| SIL | SIL | silence |

In the following sections we describe a series of experiments based on phonetic recognition which illustrate the challenges in developing speech recognition systems for children.

## 2.2. Experiments with the Italian Children's Speech Corpus

Phonetic recognition experiments were conducted using 42 held-out speakers from the FBK ChildIt corpus. For phonetic recognition we utilized the phoneme set shown in Table 1 consisting of 40 primary acoustic units (AUs). Results for phonetic recognition are presented using this 40 phoneme set as well as a reduced 33 acoustic unit set which does not take into account errors made between stressed and non-stressed vowels (e.g., "a" with "a1" and "o" with "o1").

In each experiment we utilize the phonetic sequences obtained by Viterbi alignment of the orthographic transcription of the test data as a reference phonetic transcription. 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. For each experiment described in the following sections, a 3-gram phonetic language model was estimated [13] from the resulting phonetic sequences from the phonetically aligned training data consisting of 13765 utterances.

### *2.2.1. Phonetic Recognition of Children's Speech with Adult Models*

In our first series of experiments, we wish to understand the phonetic error rate of a mismatched system (i.e., one trained on adult speech used to recognize children's speech). We also wish to quantify the error reduction which can be obtained from speaker-adaptation and normalization approaches.

For this experiment we trained adult Italian acoustic models using the FBK APASCI speech corpus [14]. 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). We use the language porting procedure outlined in Section 3.1 and estimate speaker-independent and gender-dependent models.

In order to reduce the mismatch between the adult models and children's acoustic data, we applied iterative unsupervised structural MAP linear regression (SMAPLR) using the confidence weighted phonetic recognition output [15]. The means and variances of the system Gaussians are adapted using SMAPLR after each decoding pass and used to obtain an improved phonetic recognition output. Results are shown in Table 2(a).

Previous research has also shown that vocal tract length normalization via frequency warping prior to feature extraction can assist in reducing the mismatch between children's speech and adult acoustic models. In SONIC, the frequency warping method described in [16] is implemented. The VTLN function determines the warping factor ranging between 0.88 and 1.12 per speaker such that the likelihood of the test data is maximized. Results of experiments combining SMAPLR and VTLN are summarized in Table2(b).

From Table 2(a) we can see that the initial phonetic error rate is 39.2% for a system consisting of 40 acoustic units (AU) (31.1% for 33 AUs) when adult trained acoustic models are used to recognize children's speech.

Table 2. *Children's speech Phonetic Error Rate (PER) as a function of SMAPLR adaptation iteration for Italian adult speaker independent and adult female trained acoustic models. (b) Phonetic Error Rate with SMAPLR and VTLN adaptation.*

| (a) SMAPLR Adaptation | Speaker Ind. | | 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% |

| (b) SMAPLR +VTLN | Speaker Ind. | | 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 | 29.3% | 21.8% | 27.9% | 20.3% |
| +Adapt Iter. 2 | 27.8% | 20.3% | 26.4% | 18.9 |
| +Adapt Iter. 3 | 27.1% | 19.7% | 25.9 | 18.4 |
| +Adapt Iter. 4 | 26.9% | 19.4% | 25.5 | 18.1 |
| +Adapt Iter. 5 | 26.7% | 19.3% | 25.4 | 18.0 |

As expected, adult female trained acoustic models provides some degree of improvement over speaker-independent adult models - reducing the initial phonetic error rate to 36.8% and 28.7 for 40 or 33 AUs respectively. Adaptation using SMAPLR further reduces the phonetic error rate to 28.1% and 20.7% (see Table 2(a)) for 40 or 33 AUs respectively.

Combining VTLN in the feature-space with SMAPLR in the model-space reduces the error rate to 26.7% and 19.3% (Table 2(b)). In summary, a relative error reduction of almost 32% can be achieved by combining acoustic-space adaptation (SMAPLR) and feature-space adaptation (VTLN) to adult female trained acoustic models. In the next section, we will develop acoustic models trained solely on children's speech in order to illustrate the degree of mismatch which still exists between the adapted adult models and children's speech models.

### *2.2.2. Viterbi Training of Italian Children's Speech Models*

As mentioned in Section 2.1, the language porting method of SONIC relies on an initial knowledge-based phonetic mapping between target and source language phonemes. To date, SONIC has been ported to nearly 20 languages and experience has shown that the accuracy of the initial mapping has minimal impact on the final error rate of the resulting acoustic models. In this paper, a total of 6 Viterbi alignment and acoustic model retraining passes were made to obtain final Italian children's acoustic models. In Table 3, we illustrate phonetic error rate as a function of model alignment iteration. It is clear that 6 alignment passes are sufficient to achieve system convergence. It is worth noting that the baseline children's models provide almost a 10% relative reduction in phonetic recognition error rate compared to the best adult models which have been adapted to children's speech.

### *2.2.3. Adapting Italian Children's Acoustic Models*

We extend on our baseline children's models for Italian by including iterative SMAPLR adaptation in a manner similar to that applied to the adult model experiments. Results of this experiment are shown in Table 4(a).

Table 3. *Phonetic Error Rate (PER) as a function of Viterbi alignment / model retraining pass for the FBK ChildIt Corpus. Note that Alignment Pass 0 is obtained by bootstrapping with U.S. English acoustic models while Pass 1-6 are obtained using Italian Children's models estimated from the previous Viterbi data alignments.*

| 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% |

Unlike the adult model experiments, fewer iterations of adaptation are required to achieve the lowest phonetic error rate. The acoustic adaptation applied to children's models further reduces the PER by nearly 9% relative. As in the case of recognition using adult trained models, we are interested in demonstrating recognition on children's speech where vocal tract differences between children in the training set are removed. We have thus 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 [16]. We can see from Table 4 that incorporating VTLN reduces the phonetic error rate from 21.8% to 18.7% for the 40 phonetic unit system and from 15.1% to 12.3% for the reduced 33 phonetic unit system. As anticipated, the gains from VTLN are less substantial when applied solely to children's data compared to conditions of more significant mismatch (i.e., adult model with children speech).

Table 4. *(a) Phonetic Error Rate (PER) as a function of SMAPLR adaptation iteration. (b) Phonetic Error Rate (PER) as a function of SMAPLR/VTLN adaptation iteration.*

| 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% |

Speaker Adaptive Training (SAT) attempts to remove speaker-specific characteristics from the training data in order to estimate speaker-independent acoustic model parameters. With the SONIC speech recognition system, we implement SAT by estimating a single linear feature-space transformation for each training speaker. The transform is estimated to maximize the likelihood of the training data given the VTLN normalized children's acoustic model. During testing, the VTLN warp factor is estimated along with a single Constrained MLLR (feature-space) transform prior to recognition. This final system was found reduce the PER

from 21.8% to 18.6% for the 40 unit system and from 15.1% to 12.2% for the reduced 33 unit system as shown in Table 5.

Table 5. *Phonetic Error Rate (PER) for a system combining SMAPLR, VTLN and Speaker Adaptive Trained (SAT) children's speech models.*

| 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% |

## 3. Discussion

Using the FBK ChildIt, a phonetic recognition error rate of 21.8% was achieved for first-pass recognition using a phonetic inventory of 40 units. Using a collapsed representation of 33 units, a baseline error rate of 15.1% was demonstrated. By utilizing a combination of Vocal Tract Length Normalization (VTLN), Structural MAP Linear Regression (SMAPLR); and Speaker Adaptive Training, it was demonstrated that the phonetic recognition error rate could be reduced to 18.6% for the 40 unit system and 12.2% for the reduced 33 unit system. While the error rate for the current children's system is comparable with other results reported on that corpus (compare to 22.7% for a similar 28 unit system in [17]), there still exists a significant performance gap for acoustic models which have been trained on adult speech but used to decode children's speech. Several means for acoustic adaptation including VTLN and SMAPLR were investigated to reduce acoustic mismatch. When both VTLN (feature-space transform) and SMAPLR (model-space transform) are applied to the mismatched adult/child condition, the final system achieved a phonetic error rate of 26.7% for the 40 phonetic unit system and 19.3% for the reduced 33 unit system. While these methods were shown to reduce the phonetic error rate by 28%, a 30% relative performance gap between adapted adult models and well-trained children's models still remains.

## 4. Conclusions and Future Work

Developing children's speech recognition systems for new languages presents a challenging problem due to lack of data resources. In this paper, we ported the SONIC LVSR system from English to Italian and have begun to consider the problem of optimization of the speech recognition system for Italian children's speech.

While initial corpora are available for Italian children's speech research, there exist many languages which completely lack children's speech corpora. We propose that learning statistical mappings between the adult and child acoustic space using existing adult/children corpora may provide a future direction for generating children's models for such data deficient languages.

The Italian acoustic models developed were successfully integrated into the Colorado Literacy Tutor software [1] and used to enable reading tracking in Italian for children's interactive books [18].

In a future study we will consider advanced methods for generating children's acoustic models using this work to provide a baseline for comparison

## 5. Acknowledgments

Our great thanks go the whole CSLR group at Colorado University and in particular to Bryan Pellom (now at Rosetta Stone) for his invaluable help, useful suggestions and true friendhip.

## 6. References

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

[2] Cole R., van Vuuren S., Pellom B., et al. 2003. "Perceptive Animated Interfaces: First Steps Toward a New Paradigm for Human Computer Interaction", in Proc. of the IEEE, vol. 91, no. 9, pp. 1391-1405, Sept., 2003

[3] Pellom B. 2001. "SONIC: The University of Colorado Continuous Speech Recognizer", Technical Report TR-CSLR-2001-01, University of Colorado, USA, 2001.

[4] Pellom B. and Hacioglu K. 2003. "Recent Improvements in the CU SONIC ASR System for Noisy Speech: The SPINE Task", Proc. ICASSP, Hong Kong, 2003.

[5] Hagen A., Pellom B., and Cole R. 2003. "Children's Speech Recognition with Application to Interactive Books and Tutors", Proc. ASRU, St. Thomas, USA, 2003.

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

[7] 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.

[8] Lee K., Hagen A., Romanyshyn N., Martin S., and Pellom B. 2004. "Analysis and Detection of Reading Miscues for Interactive Literacy Tutors", Proc. 20th Int. Conf.on Computational Linguistics (Coling), Geneva, CH, 2004.

[9] 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

[10] Cosi P., Pellom B. 2005. "Italian Children's Speech Recognition For Advanced Interactive Literacy Tutors", in CD-Rom Proceedings INTERSPEECH 2005, Lisbon, Portugal, 2005, pp. 2201-2204.

[11] Shobaki K., Hosom J.P., and Cole R. 2000. "The OGI Kids' Speech Corpus and Recognizers", Proc. ICSLP, Beijing, China, 2000.

[12] Gerosa M., Giuliani D. and Brugnara F. 2007. "Acoustic Variability and automatic recognition of children's speech", Speech Communication, Vol. 49, 2007, Proc. ICASSP, Hong Kong, 2003.

[13] Clarkson P.R. and Rosenfeld R. 1997. "Statistical Language Modeling Using the CMU-Cambridge Toolkit", Proc. Eurospeech, Rhodes, Greece, 1997.

[14] http://www.elda.org/catalogue/en/speech/S0039.html

[15] Siohan O., Myrvoll T., and Lee C.H. 2002. "Structural Maximum a Posteriori Linear Regression for Fast HMM Adaptation", Computer, Speech and Language, 16, 5-24, Jan, 2002.

[16] Welling L., Kanthak S., Ney H. 1999. "Improved Methods for Vocal Tract Length Normalization", Proc. ICASSP, Phoenix Arizona, 1999.

[17] Giuliani D. and Gerosa M. 2003. "Investigating Recognition of Children's Speech", Proc. ICASSP, Hong Kong, 2003.Smith, J. O. and Abel, J. S., "Bark and ERB Bilinear Transforms", IEEE Trans. Speech and Audio Proc., 7(6):697-708, 1999.

[18] Cosi P., Delmonte R., Biscetti S., Cole R., Pellom B. and van Vuuren S. 2004. "Italian Literacy Tutor: tools and technologies for individuals with cognitive disabilities", Proc. InSTIL/ICALL Symposium, Venice, Italy.