Comparing SPHINX vs. SONIC Italian Children Speech Recognition Systems

Mauro Nicolao[1,2], Piero Cosi[2]
[1] Speech and Hearing Research Group, University of Sheffield, United Kingdom
[2] Istituto di Scienze e Tecnologie della Cognizione, Consiglio Nazionale delle Ricerche, Padova, Italia
mauro.nicolao@gmail.com, piero.cosi@pd.istc.cnr.it

1. ABSTRACT

Our previous experiences have showed that both CSLR SONIC and CMU SPHINX are two versatile and powerful tools for Automatic Speech Recognition (ASR). Encouraged by the good results we had had, these two sets of tools have been compared in another important challenge of ASR: the recognition of children's speech.

In this work, SPHINX has been used to build from scratch a recognizer for Italian children's speech and the results have been compared to those obtained with SONIC both in previous and in the new experiments, which has been performed in order to uniform experimental conditions between the two different systems.

This report describes the training process and the evaluation methodology regarding a speaker-independent phonetic-recognition task. First, we briefly describe the system architectures and their differences, and then we analyze the task, the corpus and the techniques adopted to face the problem. The scores of multiple recognition tests in terms of Phonetic Error Rate (PER) and an analysis on the usability and on differences of the two systems are shown in the final discussion.

SONIC turned out to have the best overall performance and it obtained a minimum PER of 12.4% with VTLN and SMAPLR adaptation. SPHINX has been the easiest system to train and test and its performances (PER of 17.2% with comparable adaptations) were only a few percentage points far from those in SONIC.

2. INTRODUCTION

During the last few years, many different Automatic Speech Recognition frameworks have been developed for research purposes. The experience that we had with some of these systems at ISTC-CNR of Padova has confirmed that CSLR SONIC[1] and CMU SPHINX[2] are two versatile and powerful tools to build a state-of-the-art ASR. Thus, we decided to benchmark them in different contexts (Cosi et alii, 2007; Cosi & Nicolao, 2009).

In our former work, we performed a simple task for continuous clean-speech recognition of the Arabic Language with a 1k-word vocabulary. Although the task was quite simple, the result was encouraging (in terms of Word Error Rate, WER, SONIC scored 1.9% and SPHINX did 1.3%). The simplicity, with which we were able to configure the systems

[1] The SONIC speech recognition system is available for research use from the University of Colorado (http://cslr.colorado.edu)
[2] The SPHINX system (training software + decoder) is available at (http://cmusphinx.sourceforge.net/)

for such a phonetic- and spelling-complicated language, was the most interesting feature of these experiments. In our more recent work, these two systems have been tested in an evaluation campaign on the automatic recognition of connected digit for Italian language, named EVALITA[3]. The results were also extremely good (the SONIC word error rate was 2.7% and the SPHINX one was 4.5%) and both systems yielded the best performances among the other competitors.

Present work goes further applying the SPHINX system to the same children's speech recognition task where SONIC had good results in our past experiences (Cosi, 2009; Cosi & Pellom, 2005; Cosi, 2008) in order to compare the performance of the two systems.

3. SYSTEM DESCRIPTIONS

SONIC and SPHINX ASR systems have been considered in our tests because they are easily comparable. Both are statistical-model based; they use the Hidden Markov Models (HMM) to describe the acoustic feature space and Finite State Grammars or Markov chains to model the structure of language. Even so, they have important differences.

3.1 CSLR SONIC

CSLR SONIC (Pellom & Hacioglu, 2004) is a complete toolkit for research and development of new algorithms for continuous speech recognition. The software has been developed at the Center for Spoken Language Research (CSLR) of the University of Colorado since March 2001. It allows for two modes of speech recognition: Finite-State-Grammar decoding and N-gram Language Model decoding.

Since 2005, SONIC has been no more developed and the most updated version is the 2.0-beta5. Nonetheless, it is still one of the most advanced and easy-to-tune toolkit available even in comparison with a constantly developed ASR such as SPHINX. Unfortunately, since its code is not open source, only the functions and the API can be used without neither modifying nor using them in different applications.

SONIC is based on a Continuous Density Hidden Markov Model (CD-HMM) technology and it also incorporates standard speaker adaptation techniques in the recognition process, such as unsupervised Maximum Likelihood Linear Regression (MLLR) and Structural MAP Linear Regression (SMAPLR). It allows also for some adaptation methods in the training process, such as Vocal Tract Length Normalization (VTLN), cepstral mean and variance normalization, and Speaker-Adaptive Training (SAT). A good description of all these SONIC techniques can be found in (Cosi, 2009).

The software version used in our experiments was actually the older 2003 2.0-beta3, because it incorporates some powerful characteristics, no longer present in the following versions, i.e. an interesting noise-robust type of acoustic features. These are known as Perceptual Minimum Variance Distortionless Response (PMVDR) cepstral coefficients and are also well described in (Cosi, 2009). The PMVDR cepstral coefficients provide an improved accuracy over the traditional MFCC parameters by better tracking the upper envelope of the speech spectrum. Our previous experience (Cosi & Nicolao, 2009) showed a better robustness of these features in noisy environment in comparison to MFCCs.

The SONIC Acoustic Models (AMs) are built with decision-tree state-clustered HMMs with associated gamma probability density functions to model state-durations. The HMM

[3] Evaluation of NLP and Speech Tools for Italian (http://evalita.fbk.eu/)

topology has only a fixed 3-state configuration and each state can be modelled with variable number of multivariate mixture Gaussian distributions. The HMM training process consists of the alignment of the training audio material followed by an Expectation-Maximization (EM) step in which HMM parameters are tuned. Means, covariances and mixture weights are estimated in the maximum likelihood sense. The training process iterates data alignment and model estimation several times in order to gradually achieve adequate parameter estimation.

3.2 CMU SPHINX

The CMU SPHINX system, whose basic characteristics can be found in (Lee et alii, 1990), is an open-source project, developed by Carnegie Mellon University (CMU) of Pittsburgh, which provides a complete set of functions to develop complex Automatic Speech Recognition systems. There is a very reliable set of functions for the training of the HMMs and several different choices available for the large vocabulary, phoneme recognition, and N-best decoding. There is a version to be used in embedded solution (Pocket-SPHINX), a standard versatile C-version (SPHINX-3) and the Java version (SPHINX-4) suitable for web applications. Among these decoding versions, SPHINX-3, has been chosen because it is C-based and it has fitted with our testing framework better than the Java version, with no loss of performance at all.

The training functions are shared by every decoding system and are grouped in the so-called SPHINX-Train package. As in SONIC, the training process consists in iterations of alignments and acoustic-model parameter estimation. However, a difference between the two systems can be found in the model used for the preliminary phonetic alignment. SPHINX can start from raw AMs, estimated from a uniformly spaced audio segmentation, while SONIC always needs previously trained AMs. Several loops of probability-density-function (pdf) estimation with Baum-Welch algorithm and of re-alignment between training audio and transcription are performed. Models can be computed either for isolated phonemes (Context Independent, CI) or for each tri-phones found in the training data (Context Dependent, CD). Instead of the SONIC Gamma pdf, SPHINX uses Gaussian duration models. While the number of HMM states in the training tools is potentially unlimited, the recognition software has limited the topologies to 3 or 5 states and the Language Model (LM) structure to 3-grams.

The decoding process is a conventional Viterbi search algorithm through the lattice created with the output scores of HMM and some beam-search heuristics. It uses a lexical-tree-search structure, too, in order to prune the state transitions. The MLLR, VTLN and MAP adaptation methods are also implemented in this software. These techniques are substantially the same of those in SONIC and the results confirm that the recognition improvement is comparable.

4. EXPERIMENTAL FRAMEWORK

Italian Children's speech recognition is a very peculiar task and its difficulties were investigated several times by the ISTC-CNR research group (Cosi, 2009; Cosi & Pellom, 2005; Cosi, 2008).

4.1 Dataset

As dataset for Italian Children's speech recognition, the final release of the ChildIt children's speech corpus (Gerosa et alii, 2007) has been used. This consists of audio collected

by FBK[4] from 171 children (85 females and 86 males) aged between 7 and 13 (from grade 2 up to grade 8), who were all native speakers from regions in the north of Italy. Each child provided approximately 50-60 read sentences, which were extracted from age-appropriate literature.

Following (Gerosa et alii, 2007), to test the two systems, the corpus has been 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 that contain mispronunciation and noisy words have been excluded from the following experiments, while all other sentences, even those with annotated extra-linguistic phenomena, such as human disturbs (lip smacks, breath, laugh, cough, etc.), generic noises non-overlapping with speech have been included. The orthographic transcriptions of the prompt sentences have been used for training and the automatic phonetic transcription for the test. Actually, slightly mispronounced words, such as the interrupted ones, could still be considered by modifying their phonetization and forcing a silence tag at the end of them. This has helped to prevent a co-articulation between the interrupted token and the successive word that could lead to improper model training. This was also justified by the human speech error explanation theories, which affirm that, after the occurrence of an error in speech production, there is always a little pause (~20 ms) due to of the time spent for the interrupting and the restarting actions.

4.2 Training process

As most of the ASR systems require, a list of words with their standard phonetization (lexicon), a list of extra-text words (fillers), a list of phonemes, a list of questions for tree-clustering and the accurate transcription of each training audio file must be provided to configure both systems correctly. Phone-list and decision-tree question structure are the same for both systems and have been previously compiled by expert Italian linguists (Cosi & Hosom, 2000).

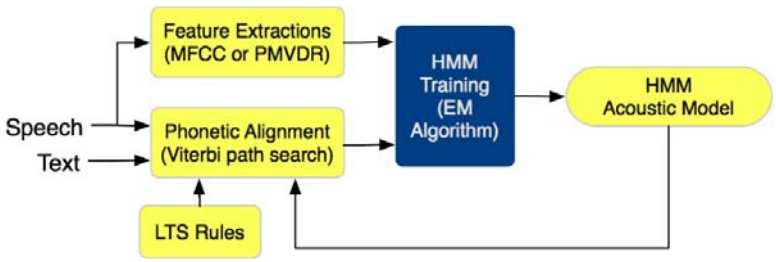

Figure 1 General scheme of both-ASR training process

Only orthographic transcriptions were available within the FBK-ChildIt corpus, thus, an automatic phonetic alignment had to be provided. This has been automatically obtained by aligning text to the audio data with a previous trained Italian AM (Cosi & Hosom, 2000) in SONIC, and with a raw AM trained by uniformly segmenting audio data in SPHINX. The latter method yields to a less accurate bootstrap, but it allows for flexibility in choosing

[4] Bruno Kessler Foundation, ex ITC-IRST (http://www.fbk.eu/)

transcriptions and phoneme list. Except from this, both training processes were quite similar and their main characteristics are summarised in Figure 1 and Table 1.

	SONIC	SPHINX
Developing centre	Centre of Spoken Language Research, University of Colorado	Carnegie Mellon University, Pittsburgh
Main reference	(Pellom & Hacioglu, 2003)	(Lee et al., 1990)
Type of ASR	Based on statistical models	
Features	39-dimension PMVDR + Δ + Δ^2	13-dimension MFCC + Δ + Δ^2 (dynamically created)
Acoustic model	Hidden Markov models with continuous density GMM	
Language model	Finite State Grammars or Markov chains	
Observation step	10 ms	
Bootstrap	unsupervised alignment with general-purpose Italian AM	uniform segmentation
Realignment	Viterbi	
HMM estimation method	Forward-backward algorithm	
CI models	none	yes (4 re-estimations)
CD models	yes (6 re-estimations)	yes (4 re-estimations)
State-clustering	yes (question-based decision tree)	
Common adaptation methods	unsupervised Maximum Likelihood Linear Regression (MLLR) and Vocal Tract Length Normalisation (VTLN), cepstral mean and variance normalisation	
Different adaptation methods	SMAPLR (Structural MAP Linear Regression), SAT (Speaker-Adaptive Training)	MAP (Maximum A-Posterior estimation)

Table 1: Main characteristics of both ASR training configurations.

Because phonetic recognition was the aim of our systems, a phoneme-level transcription, rather than a word-level one, could be given to have a simpler AM, ignoring the word-level layer. Although our work aimed to compare word-level trained models, a phoneme-level training has been also experimented. The phonetic transcription has been created with a statistical-based grapheme-to-phoneme conversion, which was trained from a 400,000-word hand-compiled Italian lexicon. The phoneme-level training has been applied exclusively to the SPHINX system. Some evidences of recognition improvement by using this method are presented.

5. RESULTS

In order to compute the score of the recognition process, Phonetic Error Rate has been used. This is defined as the sum of the deletion, substitution and insertion percentage of phonemes in the ASR outcome text with respect to a reference transcription. Ideally, a hand-labelled reference would have been preferred, because it would have been corrected at the phonetic level to take into account of children's speech pronunciation mistakes. Since this was not available, the automatic phonetic sequences obtained by a Viterbi alignment of the word-level orthographic test transcription have been used. The reference test transcrip-

tions were shared by both systems and they were created using the SONIC aligner with the general-purpose Italian model created in (Cosi & Hosom, 2000). This method has been chosen because it allowed for automatically selecting the best pronunciation for each word in the training data among the alternative choices available in the 400,000-word Italian lexicon.

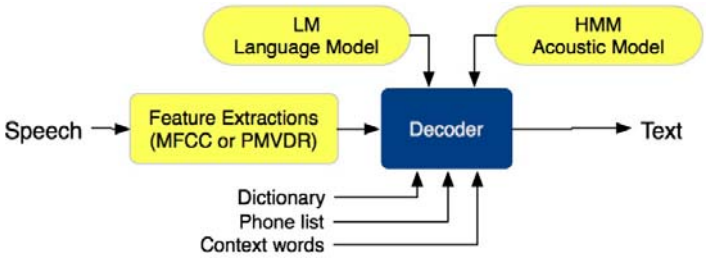

Figure 2: General scheme of both-system recognition process

Since each system has a distinct architecture, the configuration parameters are hardly comparable. It was difficult to find similarities between the two ASR systems to choose comparable configurations and to set a homogeneous test framework. To do this, we performed several experiments and finally we decided to compare the results produced by the best PER-score configuration for each test.

Only the AM performances were supposed to be evaluated, therefore we decided that the LM role, which is often too crucial in an ASR system, should have been minimized as much as possible. Thus, LM weights were set to be non-influent in the computation of the output probability of each utterance. Moreover, no symbols other than the trained phonemes were admitted. A detailed overview of the configuration for the tests is shown in the following Table 2.

	SONIC	SPHINX
Model	Tree-cluster state tied CD HMM AM	
Baseline AM	6-loop re-estimated CD	4-loop re-estimated CD
Language Model	Closed vocabulary 3-gram with null weight	
Applied adaptation methods	SMAPLR (5 loops), VTLN (warping factor 0.8-1.2)	MLLR (5 loops), VTLN (warping factor 0.8-1.2)
Model created	standard, VTLN	
Extra models	SAT	phoneme-level AM

Table 2: Testing configurations of the two systems.

Deciding when a phoneme should be considered incorrectly recognized has been another evaluation issue. In this test, the same phoneme set as in (Cosi, 2009), consisting of 40 primary Acoustic Units (AUs), has been adopted. Actually, in SPHINX, the number of primary AUs was slightly smaller, 38, because two phoneme models (/E/ and /O/) could not

be trained with the few related audio occurrences in the training data. Moreover, in unstressed position, the oppositions /e/ - /E/ and /o/ - /O/ are often neutralized in the Italian language, so it was decided to merge these couples of phonemes. Besides, the number of occurrences of /E/ and /O/ phonemes was so small in the test set that this simplification had no influence in the test results.

The acoustic differences between stressed and unstressed vowels in Italian are subtle and mostly related to their duration. Furthermore, most of the Italian people pronounce vowels according to their regional influences instead of correct pronunciation and, in children's speech production this sort of inaccuracies is even more common. For these reasons, recognition outputs have been evaluated using a 38-phoneme set as well as a reduced 33-AU set, which did not count the mistakes between stressed and unstressed vowels, and a 29-AU set with an additional phonetic simplification: i.e. geminates merged into single phoneme and reduction of /ng/ and /nf/ to /n/ (see Table 3).

Phoneme	40 AU (SONIC)	38 AU (SPHINX)	33 AU	29 AU
/i/, /o/, /u/, /k/, /t/, /tS/, /l/, /s/, /e/, /a/, /j/, /p/, /ts/, /m/, /J/, /f/, /S/, /g/, /d/, /dZ/, /r/, /z/, /w/, /b/, /dz/, /n/, /L/, /v/, SIL	present			
/E/	present	changed into /e/		
/O/	present	changed into /o/		
/ng/	present			changed into /n/
/i1/	present		changed into /i/	
/E1/	present		changed into /e/	
/o1/	present		changed into /o/	
/u1/	present		changed into /u/	
/nf/	present			changed into /n/
/e1/	present		changed into /e/	
/a1/	present		changed into /a/	
/O1/	present		changed into /o/	

Table 3. Phoneme sets (SAMPA) used for Italian Children's Speech Recognition.

The phonetic recognition experiments have been conducted using 2299 sentences from 42 held-out speakers of the FBK-ChildIt corpus. Following (Cosi, 2009), several configurations have been provided and all the efforts were made to maintain the comparability between the different configurations. All adaptations were unsupervised, which means that the reference transcriptions used to adapt the AMs were the unverified output of the adapted recognition in the previous step. The a-priori-known identities of the test speakers have been used to perform the adaptation to each speaker. Furthermore, two extra non-comparable tests have been done (one in SONIC and in one SPHINX) to check other potentialities of the systems.

As adaptation methods in SONIC, as described in (Cosi, 2009), the SMAPLR has been used; this is an iterative unsupervised structural MAP linear regression using the confidence-weighted phonetic recognition output. The means and variances of the system Gaus-

sians were adapted after each decoding pass and used to obtain an improved output. Detailed overview of the scores is displayed in Table 4(a).

SONIC with word-level training	(a) SMAPLR			(b) VTLN+SMAPLR		
	40 AU	33 AU	29 AU	40 AU	33 AU	29 AU
Baseline	21.9 %	17.2 %	15.0 %	22.2 %	17.5 %	15.3 %
Iter. 1	20.5 %	15.9 %	13.7 %	19.3 %	14.9 %	12.7 %
Iter. 2	20.0 %	15.4 %	13.2 %	19.1 %	14.7 %	12.5 %
Iter. 3	19.9 %	15.3 %	13.1 %	19.1 %	14.7 %	12.5 %
Iter. 4	19.8 %	15.2 %	13.0 %	19.0 %	14.6 %	12.4 %
Iter. 5	19.8 %	15.2 %	13.0 %	19.0 %	14.6 %	12.4 %
Baseline relative improvement	9.59 %	11.63 %	13.33 %	14.41 %	16.57 %	18.95 %

Table 4: the SONIC test: (a) PER as a function of SMAPLR adaptation iterations. (b) PER as a function of VTLN + SMAPLR adaptation iterations.

VTLN has been another applied adaptation: it is a frequency warping method of the extracted feature to compensate the peculiarities (i.e. Vocal Tract Length) of each speaker. Results of experiments combining SMAPLR and VTLN are in Table 4(b). An extra adaptation method is shown in Table 5: the Speaker Adaptive Training (SAT), not available in SPHINX, attempts to remove speaker-specific characteristics from the training data in order to estimate a set of speaker-independent acoustic models. Figure 3 summarises all the SONIC results.

SONIC with word-level training	VTLN + SAT + SMAPLR		
	40 AU	33 AU	29 AU
Baseline	21.9 %	17.2 %	15.0 %
Baseline VTLN	22.2 %	17.5 %	15.3 %
Iter. 1	18.9 %	14.7 %	12.6 %
Iter. 2	18.7 %	14.5 %	12.3 %
Iter. 3	18.8 %	14.5 %	12.3 %
Iter. 4	18.7 %	14.4 %	12.3 %
Iter. 5	18.8 %	14.4 %	12.2 %
Baseline relative improvement	14.16 %	16.28 %	18.67 %

Table 5: the SONIC extra test PER, combining SMAPLR, VTLN and the SAT children's speech models.

In SPHINX test, a MLLR has been performed instead of the SONIC SMAPLR. However, it is also an unsupervised and iterative recognition, in which the speaker-independent AMs are modified with a linear transformation in order to minimize the distance between means and variances of the speaker-independent models and those of the unknown-speaker. The detailed results are in Table 6(a).

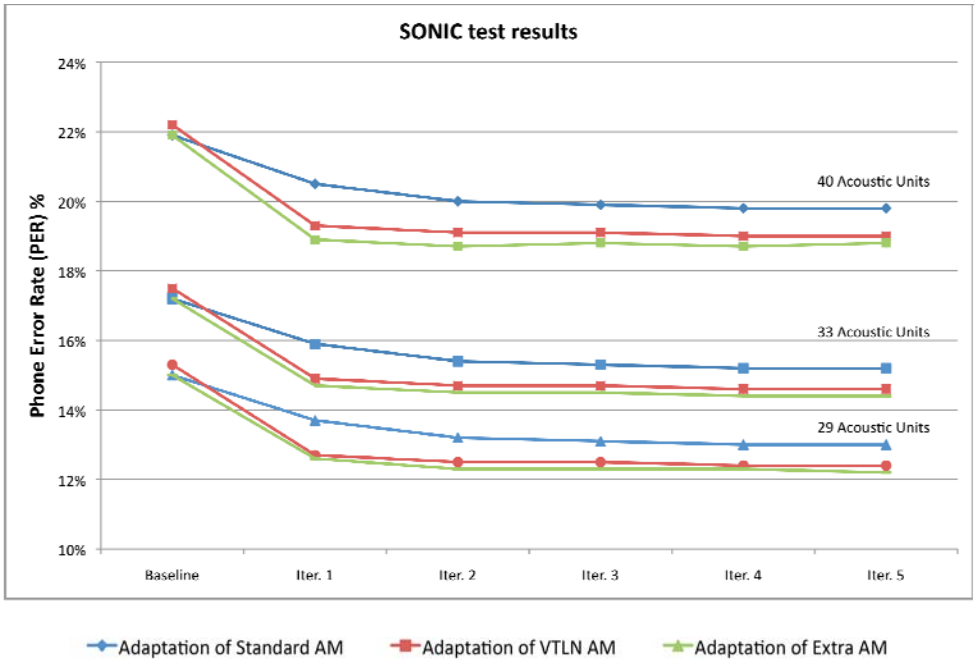

Figure 3: Plot of the SONIC test results: PER as a function of SMAPLR adaptation iterations; PER as a function of SMAPLR +VTLN adaptation iterations; PER of extra test, combining SMAPLR, VTLN and SAT children's speech models.

SPHINX with word-level training	**(a) MLLR**			**(b) VTLN+MLLR**		
	40 AU	**33 AU**	**29 AU**	**40 AU**	**33 AU**	**29 AU**
Baseline	29.8 %	22.1 %	19.2 %	29.2 %	21.5 %	18.7 %
Iter. 1	29.1 %	20.9 %	18.1 %	28.7 %	20.6 %	17.8 %
Iter. 2	28.7 %	20.5 %	17.7 %	28.5 %	20.3 %	17.5 %
Iter. 3	28.5 %	20.3 %	17.5 %	28.4 %	20.2 %	17.4 %
Iter. 4	28.3 %	20.1 %	17.3 %	28.4 %	20.2 %	17.4 %
Iter. 5	28.2 %	20.0 %	17.2 %	28.3 %	20.1 %	17.3 %
Baseline relative improvement	5.37 %	9.50 %	10.42 %	3.08 %	6.51 %	7.49 %

Table 6: the SPHINX test: (a) PER as a function of MLLR adaptation iterations. (b) PER as a function of VTLN + MLLR adaptation iterations.

In SPHINX, the VTLN method has been also applied. The related recognition scores along with a further 5-loop MLLR adaptation are in Table 6(b). Finally, the extra experiment results are displayed in Table 7. As mentioned above, a phonetically annotated transcription has been used to train the baseline AM and the same MLLR adaptation has been performed during recognition.

SPHINX phoneme-level training	VTLN + MLLR		
	40 AU	33 AU	29 AU
Baseline	28.2 %	21.2 %	18.6 %
Baseline VTLN	27.5 %	20.5 %	18.0 %
Iter. 1	26.4 %	19.4 %	17.0 %
Iter. 2	26.0 %	19.1 %	16.7 %
Iter. 3	25.8 %	18.9 %	16.6 %
Iter. 4	25.7 %	18.8 %	16.5 %
Iter. 5	25.7 %	18.8 %	16.4 %
Baseline relative improvement	8.87 %	11.32 %	11.83 %

Table 7: the SPHINX extra test: PER as a function of VTLN + MLLR adaptation iterations for a system trained with phonetic transcriptions.

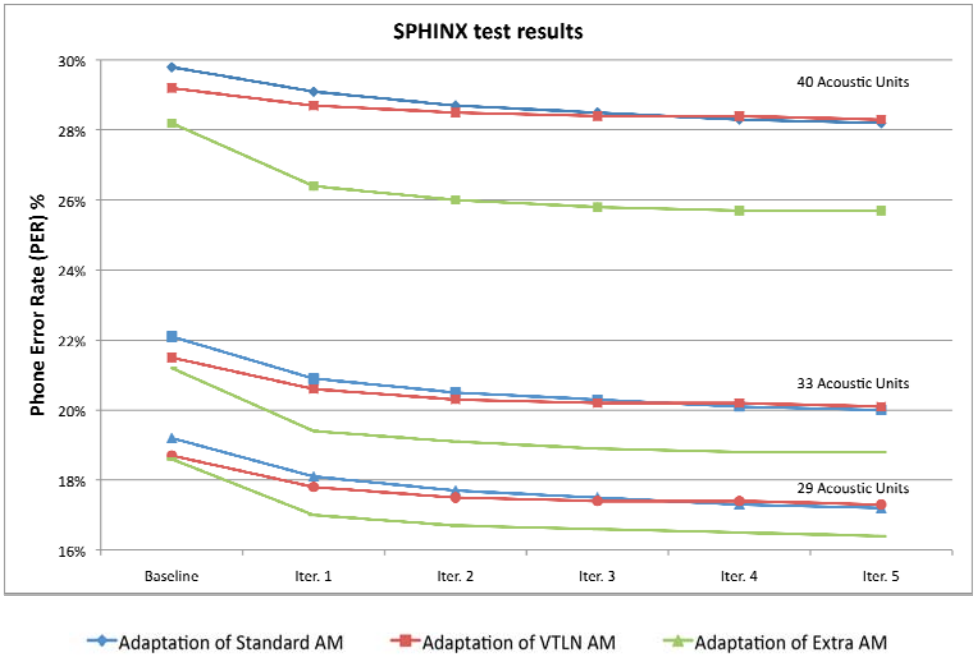

Figure 4: Plot of the SPHINX test results: PER as a function of MLLR adaptation iterations; PER as a function of MLLR +VTLN adaptation iterations; PER of extra test, VTLN + MLLR adaptation iterations for a system trained with phonetic transcriptions.

6. FINAL DISCUSSION

Comparing the scores of the word-level-trained systems in Table 8, it can be observed that, even if both had good performance, SONIC has obtained better scores with 5-8 points of percentage, on average. This gap reduces as far as the number of AUs decreases. This is

mainly because the number of mistakes is spread over a smaller number of phones, but this could also mean that SONIC was able to recognize little differences between phonemes, such as accents or durations, better than SPHINX.

	SONIC			SPHINX		
	40AU	33AU	29AU	40AU	33AU	29AU
Baseline	21.9%	17.2%	15.0%	29.8%	22.1%	19.2%
Best score	19.0%	14.6%	12.4%	28.2%	20.0%	17.2%

Table 8: Comparison between PER scores of the different configurations of SONIC and SPHINX.

	SONIC VTLN+ SAT +SMAPLR			SPHINX phoneme-level training		
	40AU	33AU	29AU	40AU	33AU	29AU
Baseline	21.9%	17.2%	15.0%	28.2%	21.2%	18.6%
Best score	18.8%	14.4%	12.2%	25.7%	18.8%	16.4%

Table 9: Comparison of the extra-test PER scores.

Considering the baseline-relative improvements obtained by applying the VTLN and the model adaptations to both systems for each AU configuration, they turn out to be between 9.6% and 18.6% in SONIC and between 3% and 10.4%. This means, that adaptation methods of both systems have been effective but SMAPLR has been slightly better than MLLR in adapting the baseline models.

However, the main cause of such difference in the absolute performance is the baseline model. SONIC model had better results due to several reasons; in particular, two factors have played a decisive role. First, the PMVDR features used to describe the signal spectra in SONIC apparently react very well to high-variable signal such as children's speech. The second reason was the bootstrap of the training process: while SONIC could take advantages of a good first segmentation, SPHINX had a raw uniformly spaced one. Some further tests are already planned in order to investigate this differences more deeply.

Concerning the extra tests, an improvement has been obtained using the SPHINX model trained with phonetic transcriptions. Since the same configuration and the same parameters of the standard test were used, the improvement (significative in the 40-AU case) has been totally related to the different type of training. The adaptation has been also more effective in this extra test and the PER relatively improves of about 3-4 points in comparison to the standard case. For what concerns the SONIC extra test, the absolute results and the adaptation effectiveness were very similar to those of the standard case.

A marginal consideration might be added as a final conclusion. Even if SONIC turns out to have the best overall performances, nonetheless, it has begun to be less attractive than SPHINX for research purpose because its libraries are a close piece of software and there is no more development and support from CSLR. Moreover, the apparently so efficient PMVDR are no more included in the standard distribution.

7. REFERENCES

Cosi, P., Hosom, J. P. (2000), High Performance General Purpose Phonetic Recognition for Italian, *Proc. ICSLP 2000*, Beijing, 527-530.

Cosi P. and Pellom B. (2005), Italian Children's Speech Recognition For Advanced Interactive Literacy Tutors, *Proc. INTERSPEECH 2005*, Lisbon, Portugal, 2201-2204.

Cosi, P., Nicolao, M., Sommavilla, G. and Tisato, G. (2007), Sviluppo di un sistema di riconoscimento per l'Arabo: problemi e soluzioni, *Proc. of AISV 2007, 4th Conference of AISV*, Reggio Calabria, Italy.

Cosi, P. (2008), Recent Advances in Sonic Italian Children's Speech Recognition for Interactive Literacy Tutors, *1st Workshop On Child, Computer and Interaction (WOCCI)*, Chania, Greece.

Cosi, P. and Nicolao, M. (2009), Connected Digits Recognition Task ISTC-CNR Comparison of Open Source Tools, *Proc. of the 11th Conference of the Italian Association for Artificial Intelligence*, Reggio Emilia, Italy.

Cosi, P. (2009), On the Development of Matched and Mismatched Italian Children's Speech Recognition Systems, *Proc INTERSPEECH 2009*, Brighton, UK, 540-543.

Gerosa M., Giuliani D. and Brugnara F. (2007), Acoustic variability and automatic recognition of children's speech, *Speech Communication*, 49, 847-860.

Lee, K. F., Hon, H. W., Reddy, R. (1990), An overview of the SPHINX speech recognition system, *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 38, 35-45.

Pellom, B., Hacioglu, K. (2004), SONIC: Technical Report TR-CSLR-2001-01, *Center for Spoken Language Research*, University of Colorado, Boulder.