Connected Digits Recognition Task: ISTC–CNR Comparison of Open Source Tools

Piero Cosi and Mauro Nicolao

Istituto di Scienze e Tecnologie della Cognizione, C.N.R.
via Martiri della libertà, 2 – 35137 Padova (ITALY)
piero.cosi@pd.istc.cnr.it, mauro.nicolao@pd.istc.cnr.it

Abstract. EVALITA is a recent initiative devoted to the evaluation of Natural Language and Speech Processing tools for Italian. In this work, the results of three open source ASR toolkits will be described. CSLU Speech Tools, CSLR SONIC, CMU SPHINX are applied on the EVALITA clean and noisy digits recognition task and this report will describe the complete evaluation methodology. CSLR SONIC has resulted to have the best performances in all the tasks and even with high specialized trainings. We think that it is mostly because of the PMVDR features used in this system. CMU SPHINX has been the easiest system to train and test and its general performances are only slightly lower than SONIC. CSLU Speech Tools is the most specialized recognition system on digit and its score stands in the middle of the others. Overall, the three systems have Word Accuracy score over 90%.

Keywords: EVALITA, CSLU Speech Tools, CSLR SONIC, CMU SPHINX, clean and noisy speech, Artificial Neural Network, Hidden Markov Models.

1 Introduction

EVALITA provides a shared framework for the different systems and approaches on separate natural language and speech processing tasks for the Italian language. In EVALITA 2009, several tasks have been evaluated but we focused only on *Speech tasks*, mainly on Connected Digits Recognition. In this task, systems are required to recognize sequences of spoken Italian digits (numbers ranging from 0 to 9). The Analysis Corpus consists of 16kHz, 16bit-PCM, mono Windows wav audio files. Two subtasks are defined: *clean speech audio* with audio recorded in clean environment and *noisy speech audio* in which audio acquired in noisy environment. The type of noise may vary from white noise to traffic, room, etc.

The evaluation process is based on Minimum Edit Distance between the transcriptions coming from the recognizer and the orthographic annotations. Accuracy will be calculated at word and phrase levels and participants which need to enrol the ASR at finer level than phrase have to provide by themselves for the annotation.

The results of the EVALITA Test are computed on the test-audio transcriptions recognized by the ASR systems with the configurations previously tuned on development files.

2 System description

The aim of our test is to compare three of the most used open source tools for the Automatic Speech Recognition. CSLU Toolkit, CSLR SONIC and CMU SPHINX were considered because promising results were obtained in the past on similar digit recognition tasks.

2.1 CSLU Toolkit

The CSLU Toolkit[1] is a comprehensive set of tools for learning about, researching and developing interactive language systems and their underlying technologies. The CSLU Toolkit has been described in several papers [1], [2] and will not be detailed here. The basic framework of the CSLU Toolkit is represented by an hybrid Hidden Markov Model (HMM) and Artificial Neural Network (ANN) architecture in which the phonetic likelihoods are estimated using a neural network instead of a mixture of Gaussians, which has the advantage of not requiring assumptions about the distribution or independence of the input data, of easily performing discriminative training and of splitting each phoneme into states that are dependent on the left or right context, or are context independent [3]. As for feature extraction MFCC, MFCC+Δ, MFCC+Δ +Δ^2 and PLP+MFCC, added by Cepstral Mean Subtraction (CMS), and RASTA were compared.

Neural-network training was done with standard back-propagation on a fully connected feed-forward network. The training data were searched to find all the vectors of each category in the automatically-labelled training section. The neural network was trained using the back-propagation method to recognize each context-dependent category in the output layer. Each training waveform was then recognized using the best obtained network (Baseline), with the result constrained to be the correct sequence of digits. This process, called *Forced-Alignment* (FA), was used to generate time-aligned category labels. These FA category labels were then used in a second cycle of training and evaluation was repeated to determine the new best network, which was finally evaluated on the development data.

In order to improve the recognition results, the *Forward-Backward* (FB) training strategy was recurrently applied (three times) [4]. Like most of the other hybrid systems, the neural network in this system is used as a state emission probability estimator. A three-layer fully connected neural network can be conceived, with the same configuration as that of the baseline and forced-aligned neural networks and the same output categories. Unlike most of the existing hybrid systems which do not explicitly train the within-phone relative likelihoods, this new hybrid trains the within-phone models to probability estimates obtained from the forward-backward algorithm, rather than binary targets. To start FB training an initial binary-target neural network is required. For this initial network, the best network resulting from forced-alignment training (FA) was used. Then the Forward-Backward re-estimation algorithm was used to regenerate the targets for the training utterances. The re-

[1] The CSLU Toolkit is available through the CSLU OGI Web site: http://cslu.cse.ogi.edu/toolkit/.

estimation was implemented in an embedded form, which concatenates the phone models in the input utterance into a big model and re-estimates the parameters based on the whole input utterance. The networks would be trained using the standard stochastic back-propagation algorithm, with mean-square-error as the cost function.

2.2 CSLR SONIC

SONIC[2] is a complete toolkit for research and development of new algorithms for continuous speech recognition. The software has been under development at CSLR since March of 2001 at the University of Colorado. The current implementation allows for two modes of speech recognition: Keyword and Finite State Grammar decoding and N-gram language-model decoding.

SONIC is based on Continuous Density Hidden Markov Model (CDHMM) technology and it incorporates speaker adaptation and normalization methods such as Maximum Likelihood Linear Regression (MLLR), Vocal Tract Length Normalization (VTLN), and cepstral mean and variance normalization.

In SONIC version 2.0-beta3, CSLR has adopted an acoustic feature representation known as Perceptual Minimum Variance Distortionless Response (PMVDR) cepstral coefficients [5]. PMVDR cepstral coefficients provide improved accuracy over traditional MFCC parameters by better tracking the upper envelope of the speech spectrum. Unlike MFCC parameters, PMVDRs do not require an explicit filter-bank analysis of the speech signal (see [6] and [7] for details).

The acoustic models consists of decision-tree state-clustered HMMs with associated gamma probability density functions to model state-durations. The acoustic models have a fixed 3-state topology. Each HMM state can be modelled with variable number of multivariate mixture Gaussian distributions.

The training process consists of first performing state-based alignment of the training audio followed by an expectation-maximization (EM) step in which decision-tree state-clustered HMMs are estimated. Acoustic model parameters (means, covariances, and mixture weights) are estimated in the maximum likelihood sense. The training process can be iterated between alignment of data and model estimation to gradually achieve adequate parameter estimation.

2.3 CMU SPHINX

SPHINX[3] system [8] is a open-source project which provides a complete set of functions to develop complex Automatic Speech Recognition systems. This software has been developed by Carnegie Mellon University at Pittsburgh. It includes both an acoustic *trainer* and various *decoders*, for text recognition, phoneme recognition, N-best list generation and more.

SPHINX training is an iterative sequence of alignments and AM-estimations. It starts from an audio segmentation aligned to training-data transcriptions and it

[2] SONIC was available through the CSLR Web site: http://sonic.colorado.edu/sonic/download/index.html.

[3] The SPHINX system is available at http://cmuSPHINX.sourceforge.net/html/cmuSPHINX.php.

estimates a raw first AM from them. This is the starting point of the following loops of Baum-Welch probability density functions estimation and transcription alignment. Models can be computed either for each phoneme (Contest Independent, CI) or, considering phoneme context (Contest Dependent, CD). SPHINX acoustic models are trained over MFCC + $\Delta + \Delta^2$ feature vectors.

While the training process is unique, in the decoding step different versions of the recognizer can be used. We adopted SPHINX-3, which is a C-based state-of-the–art large-vocabulary continuous-model ASR, in order to better merge SPHINX tools with our test framework. It is limited to 3 or 5-state left-to-right HMM topologies and to a bigram or trigram language model. The decoder is based on the conventional *Viterbi search* algorithm and *beam search* heuristics. It uses a *lexical-tree* search structure, too, in order to prune the state transitions.

As the other systems, it produces a single best recognition result (or hypothesis) for each utterance processed which is a linear word sequence.

3 Experiments

In this paragraph, we are going to describe the structures and the parameters of our experiments.

3.1 Evalita Data

EVALITA data is constituted by clean and noisy digits audio file sets. Exclusively the ten Italian digits are in the audio files. The ASR pronunciation lexicon can be the same for the three systems. It has the 10 different words shown above, plus 3 special fillers: begin and end sentence marker and silence identifier. The word phonetization derives from the SAMPA transcription: 0 [dz E r o], 1 [u n o], 2 [d u e], 3 [t r E], 4 [k w a t r o], 5 [tS i n k w e], 6 [s E I], 7 [s E t e], 8 [O t o], 9 [n O v e]. Multi pronounce entries can be produced to model regional pronunciation differences. EVALITA files are divided in three sub sets: train, development and test as shown in Table 1.

Table 1: Sub set description.

Sub Set	Clean Audio Files	Noisy Audio Files	Clean Digit Sequences	Noisy Digit Sequences
Training	3144	2204	10129	7376
Development	216	299	1629	1941
Test	365	605	2361	4036

3.2 Experimental Framework

We have trained all our systems on the EVALITA training data, we have tuned them to get the best results on the EVALITA development data and finally we have performed recognition on the EVALITA test set.

We have decided to analyze clean and noisy data both separately and together by training different acoustic models for each type of audio data. Thus, our framework for each ASR system consists of three recognition experiments with different acoustic models: all-training-file AM, only-noisy-training-file AM, only-clean-training-file AM.

Finding similarity among the three ASR systems in order to choose comparable configurations was one of the main difficulties to set an homogeneous test framework. Every system has a completely distinct architecture and consequently the configuration parameters are hardly comparable. We did several experiments and finally we decided to compare results produced by the best WA-score configurations.

More difficulties came from Language Model (LM) whose role is crucial in ASR system. CSLU toolkit admits only a Finite State Grammar to model the possible utterances. A simple grammar [<any> (<digit> [silence]) + <any>] allowing any digit sequence in any order, with optional silence between digits, was considered.

To compare SONIC and SPHINX results with CSLU, we have performed SONIC and SPHINX recognitions with LM weight set to 0. This should simulate complete independence between connected digits but using words as basic recognition unit.

Within the EVALITA framework only the orthographic transcriptions are available so one of our previously-created general-purpose recognizer [9] has been used to create the phonetically aligned transcriptions needed from CSLU and SONIC systems to start the training.

In CSLU toolkit, then, a three-layer fully connected feed-forward network was trained to estimate, at every frame, the probability of 98 context-dependent phonetic categories. These categories were created by splitting each Acoustic Unit (AU), into one, two, or three parts, depending on the length of the AU and how much the AU was thought to be influenced by co-articulatory effects. *Silence* and *closure* are 1-part units, *vowels* are 3-part units, *unvoiced plosive* is 1-part right dependent unit, *voiced plosive*, *affricate*, *fricative*, *nasal*, *liquid retroflex* and *glide* are all 2-part units. A hundred iterations ware done and the best network iteration (*baseline network* - B) was determined by evaluation on the EVALITA clean and noisy digits development sets respectively. After a comparison among the CSLU system driven by different feature types, we have found that 13-coefficient PLP plus 13-coefficient MFCC with CMS obtained the best score.

As for SONIC, 12 PMVDR cepstral parameters were retained and augmented with normalized log frame energy plus the first and second differences of the features. A final 39-dimensional feature vector is computed, once every 10 ms. Then, the model developed in [9] was inserted in the first alignment step to provide a good segmentation to start from and a first acoustic-model estimation was computed. At the end of further eight loops of phonetic alignment and acoustic model re-estimation, the final AM is considered well trained.

In SPHINX training no previously developed AM was applied and a simple uniform segmentation was chosen as starting point. After raw first-AM estimation, four loops of re-alignment and contest-independent AM re-estimations were done. The last CI trained model was employed to create a minimum-error segmentation and train contest-dependent AMs. First an all-state (untied) AM was computed, and then four loops of CD state-tied segmentation–re-estimation were done.

4 Results obtained

In order to be able to best-tune the different system performances, according to the EVALITA evaluation rules, we used a tool in the NIST SCTK Scoring Toolkit[4], a NIST SCLITE software. Development performances have been computed by using EVALITA development transcriptions as reference, but, as for Test results, some missed and misspelled words have been added to EVALITA official test transcriptions.

In Table 2 and Table 3 the results for the *clean*, *noisy*, and *clean plus noisy* experiments for CSLU Toolkit are summarized. It shows, as expected, quite good performance with clean digit sequences, and also quite promising results in the noisy and clean + noisy case. Finally, the final test on the EVALITA clean and noisy digits test-sets is executed with the best obtained network (FB1).

Table 2: CSLU Toolkit ASR results in terms of Word Accuracy for Development set. IA means Initial Alignment, FA is Force Alignment, FBn is the n-th loop of Forward Backward process.

Development WA %	IA	FA	FB1	FB2	FB3
Clean AM on clean	99,82	99,75	**99,94**	99,82	99,75
Noisy AM on noisy	90,15	90,93	**92,11**	91,75	91,49
Full AM on clean+noisy	93,86	94,12	**94,28**	**94,28**	94,2

Table 3: CSLU Toolkit ASR results in terms of Word Accuracy and Sentence Accuracy for Test set. *Full AM* means recognition made by clean-plus-noisy audio-file trained AM; *Clean AM* means recognition made by clean-audio-file trained AM; *Noisy AM* means recognition made by noisy-audio-file trained AM.

Test FB1	WA %	SA %
Clean AM on clean	99,10	94,80
Noisy AM on noisy	94,00	82,00
Full AM on clean + noisy	95,00	87,20

Concerning the Development data tuning of SONIC and SPHINX, we maximized the WA score and test recognition was performed with the best-WA configuration. In the following tables, Table 4, Table 5, Table 6 and Table 7, the results for the clean, noisy, and clean + noisy experiments for both SONIC and SPHINX are summarized. Concerning the Language Model, all the configurations have a null LM weight. This assumption doesn't limit the ASR performances, indeed, in the connected digit task, the result scores with not-zero LM weight were lower than the corresponding with zero LM weight, because there is complete independence between spoken units while using words as basic recognition unit.

[4] The NIST SCLITE software is available at the website: http://www.itl.nist.gov/iad/mig/tools/

Table 4: SONIC ASR results in terms of Word Accuracy for Development set.

Development WA %	Full AM	Clean AM	Noisy AM
clean	99,70	**99,80**	**99,70**
noisy	94,20	89,90	**94,80**
clean + noisy	96,71	94,42	**97,04**

Table 5: SONIC ASR results in terms of Word Accuracy and Sentence Accuracy for Test set.

Test	WA %	SA %
Clean AM on clean	99,60	97,30
Noisy AM on noisy	96,30	87,90
Full AM on clean + noisy	97,30	90,60

Table 6: SPHINX ASR results in terms of Word Accuracy for Development set.

Development WA %	Full AM	Clean AM	Noisy AM
clean	**99,40**	**99,40**	98,80
noisy	**93,30**	78,70	92,60
clean + noisy	**96,10**	88,31	95,43

Table 7: SPHINX ASR results in terms of Word Accuracy and Sentence Accuracy for Test set.

Test	WA %	SA %
Clean AM on clean	98,90	94,50
Noisy AM on noisy	91,70	72,70
Full AM on clean + noisy	95,50	86,00

5 Discussion about results

Three of the most used open source ASR tools were considered in this work, i.e. CSLU Toolkit, SONIC, and SPHINX, because promising results were obtained in the past on similar digit recognition tasks.

Beyond the fact that finding similarity among the three ASR systems was one of the main difficulties, an homogeneous and unique test framework for comparing different Italian ASR systems was quite possible and effective if 3-gram LM weight is set to 0 and the results produced by the best WA-score configuration were compared for each system.

CSLU Toolkit is good in recognizing clean digit sequences, but it is not so good at recognizing clean-plus-noisy audio. SONIC is the best system in all situations and we believe this is mainly due to the adoption of the PMVDR features. SPHINX is quite more sensible to AM specialization than other systems and clean models can not recognize noisy speech with high performance.

Table 8: Summary of Test Recognition Results.

Test	CSLR		SONIC		SPHINX	
	WA %	SA %	WA %	SA %	WA %	SA %
Clean AM on clean	99,10	94,80	**99,60**	**97,30**	98,90	94,50
Noisy AM on noisy	94,00	82,00	**96,30**	**87,90**	91,70	72,70
Full AM on clean + noisy	95,00	87,20	**97,30**	**90,60**	95,50	86,00

Finally we should conclude that the EVALITA campaign was quite effective in forcing various Italian research groups to focus on similar recognition tasks working on common data thus comparing and improving various different recognition methodologies and strategies, and we hope more complex task and data will be exploited in the future.

Acknowledgment. Our great thanks go the whole CSLU group at OGI Portland Oregon, and in particular to John Paul Hosom and Johan Shalkwyk for their tremendous patience and help in making us playing with the CSLU Speech Toolkit, and also to the whole CSLR group at Colorado University and in particular to Bryan Pellom (now at Rosetta Stone) for his invaluable help and useful suggestions in developing SONIC experiments, and for their true friendship.

References

1. Sutton S., Cole R.A., de Villiers J., Schalkwyk J., Vermeulen P., Macon M., Yan Y., Kaiser E., Rundle B., Shobaki K., Hosom J.P., Kain A., Wouters J., Massaro D., Cohen M.: Universal Speech Tools: the CSLU Toolkit. In: Proceedings ICSLP 98, vol. 7, pp. 3221--3224. Sydney, Australia (1998)
2. Cole R.A.: Tools for research and education in speech science. In: Proceedings ICPhS99, vol. 2, pp. 1277-1280. San Francisco, CA (1999)
3. Bourlard H.: Towards Increasing Speech Recognition Error Rates. In: Proceedings EUROSPEECH 95, vol. 2, pp. 883-894. Madrid, Spain (1995)
4. Yan, Y., Fanty, M., Cole, R.: Speech Recognition Using Neural Networks with Forward-Backward Probability Generated Targets. In: Proceedings ICASSP97, vol. 4, pp. 3241--3244 (1997)
5. Yapanel, U.H, Hansen, J.H.L.: A new perspective on Feature Extraction for Robust In-vehicle Speech Recognition. In: Proceedings of Eurospeech'03. Geneva, Switzerland (2003)
6. Pellom, B., Hacioglu, K., SONIC: Technical Report TR-CSLR-2001-01, Center for Spoken Language Research. University of Colorado, Boulder (2004)
7. Murthi, M.N., Rao, B.D.: MVDR Based All-Pole Models for Spectral Coding of Speech. In: Proceedings of ICASSP 99. Phoenix (1999)
8. Lee, K.F., Hon, H.W., Reddy, R.: An overview of the SPHINX speech recognition system. IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 38, no. 1., pp. 35--45 (1990)
9. Cosi, P., Hosom, J.P.: High Performance “General Purpose” Phonetic Recognition for Italian. In: Proceedings of ICSLP 2000, International Conference on Spoken Language Processing, vol. II, pp. 527--530. Beijing, China (2000)