HIGH PERFORMANCE ITALIAN CONTINUOUS "DIGIT" RECOGNITION

Piero Cosi , John-Paul Hosom** and FabioTesser****

*Istituto di Fonetica e Dialettologia -- C. N. R., Via G. Anghinoni, 10 - 35121 Padova ITALY
email: cosi@csrf.pd.cnr.it

**Center for Spoken Language Understanding (CSLU-OGI)
Oregon Graduate Institute (OGI), P.O. Box 91000, Portland, Oregon 97291 USA
email: hosom@cse.ogi.edu

***Università di Padova – Dipartimento di Elettronica e Informatica
Via Gradenigo 6/a, 35131 Padova, ITALY
e-mail: tesser@dei.unipd.it

ABSTRACT

The development of a speaker independent connected "digits" recognizer for Italian is described. The CSLU Speech Toolkit was used to develop and implement the system which is based on an hybrid ANN/HMM architecture. The recognizer is trained on context-dependent categories to account for coarticulatory variation. Various front-end processing was compared and, when the best features (MFCC with CMS + Δ) were considered, there was a 98.68% word recognition accuracy (90.76% sentence recognition accuracy) on a test set of the FIELD continuous digits recognition task.

1. INTRODUCTION

The "digits" small-vocabulary task, with ten digits from "zero" through "nine", is important for many telephone-based applications, such as computer-assisted long-distance dialing, credit-card billing and, in general, for all applications where data input by speech is necessary. Moreover this task requires extremely high accuracy, and focuses research on acoustic-level processing.

The aim of this work was that of investigating mostly the effects of the feature set in order to optimize the Italian digit recognition accuracy over the telephone channel. Various combinations of features, such as PLP [1] and MFC coefficients [2], together with two normalization procedures, such as RASTA [3] and Cepstral Mean Subtraction [4], were investigated.

2. RECOGNITION FRAMEWORK

The recognizer being described in this work was developed and implemented by the use of the CSLU Speech Toolkit [5] freely available through the CSLU OGI Web site [6]. The basic framework considered for recognition was that corresponding to an hybrid ANN/HMM architecture [7], [8]. The major difference between this framework and standard HMM systems is that the phonetic likelihoods are estimated using a neural network instead of a mixture of gaussians. A second difference is in the type context-dependent units. Whereas standard HMMs train on the context of the preceding and following phonemes, our system splits each phoneme into states that are dependent on the left or right context, or are context independent.

3. DATA

Three corpora have been used in this work in order to train, develop and test the telephone-channel digits recognition system: FIELD, PHONE [9], [10] and PANDA [11], [12]. The FIELD corpus contains telephone numbers that were collected as part of a semi-automated collect-call service, and the PHONE corpus contains random digits strings obtained from cooperative but naive speakers and has a large number of hesitations, breath noise, and other "spontaneous speech phenomena". The speech material contained in the PANDA corpus belongs, instead, to a "credit card" domain; it corresponds to various credit-card-like digit strings pronounced by more than 1000 speakers. The speech material was divided into training, development and test sub-sets.

4. EXPERIMENT

The "digits" recognizer was trained on context-dependent categories to account for coarticulatory variations and recognizes any connected sequence of the 10 Italian digits:

0 [dz E r o], 1 [u n o], 2 [d u e], 3 [t r E], 4 [k w a tt r o], 5 [tS i n k w e], 6 [s E I], 7 [s E tt e], 8 [O tt o], 9 [n O v e].

In particular the simple grammar [<any> (<digit> [silence])+ <any>] allowing any digit sequence in any order, with optional silence between digits, was considered.

4.1 Acoustic units

A three-layer fully connected feed-forward network with 130 inputs and 200 nodes in the single hidden layer, was trained to estimate, at every frame, the probability of 116 context-dependent phonetic categories. These categories were created by splitting each Acoustic Unit (AU), as illustrated in Table 1 and 2, into one, two, or three parts, depending on the length of the AU and how much the AU was thought to be influenced by coarticulatory effects. AU states were trained for different preceding and following phonetic contexts, and some phonetic contexts were grouped together to form a broad-context grouping. The broad-context groupings were done based on acoustic-phonetic knowledge.

Acoustic Units	Parts	Description
.pau @eh @br	1	silence
i e E a O o u	3	vowel
tcl kcl	1	closure
t k	r*	unvoiced plosive
d	2	voiced plosive
dz tS	2	affricate
s v	2	fricative
n	2	nasal
r	2	liquid retroflex
w	2	glide

Table 1. Acoustic units (SAMPA [13], except closures) and number of parts to split each unit into for the Italian "digits" lexicon recognizer (* r means "right dependent unit").

Group	Acoustic units in group	Description
$sil	.pau, .grb @br	silence
$pld	d t tcl	dental plosive
$alv	dz s	alveolar
$lab	v	labial
$pal	tS	palatal
$ret	r	retroflex
$nas	n	nasal
$vel	k kcl	velar
$bck	u o O w	back vowel/glide
$mid	a E	mid vowel
$frn	i, e	front vowel

Table 2. Groupings of acoustic units into clusters of similar units, for the Italian digits task.

4.2 Feature extraction

As for feature extraction, various combination of MFCC and PLP coefficients (with and without CMS and RASTA processing), plus their delta or delta-delta values were compared. They were continuously computed with a 10-msec frame rate. The input to the network consisted of the features for the frame to be classified, as well as the features for frames at -60, -30, 30, and 60 msec relative to the frame to be classified. In the case of 12 MFCC coefficients plus the energy plus their delta values the network consisted of 130 input nodes.

4.3 Training strategy

Neural-network training was done with standard back-propagation on a fully connected feed-forward network. The training was adjusted to use the negative penalty modification proposed by Wei and van Vuuren [14]. With this method, the non-uniform distribution of context-dependent classes, that is dependent on the order of words in the training database, is compensated for by flattening the class priors of infrequently occurring classes. This compensation allows better modeling for an utterance in which the order of the words can not be predicted.

4.4 Duration constraints

Transition probabilities were set to be all equally likely, so that no assumptions were made about the a priori likelihood of one category following another category. In order to make use of a priori information about phonetic durations, and to minimize the insertion of very short words, the search was constrained by specifying minimum duration values for each category, where the minimum value for a category was computed as the value at the second percentile of all duration values. During the search, hypothesized category durations less than the minimum value were penalized by a value proportional to the difference between the minimum duration and the proposed duration. The grammar allowed any phoneme in any order, with optional silence between phonemes.

4.5 "Baseline"

The "baseline" system was trained with part of the FIELD corpus (38%) corresponding to 352 digit sequences (3307 digits), and the whole PHONE corpus (2241 digit sequences, 9131 digits). Moreover, 13% of the FIELD corpus was used for the development (127 digit sequences, 1086 digits) and the remaining 49% (488 digit sequences, 4614 digits) was used for the test. In summary, 12438 hand-labeled digits were used for training, 1086 were considered for the development and 4614 for the test phase. The training data were searched to find all the vectors of each category in the hand-labeled training section. The neural network was trained using the back-propagation method to recognize each context-dependent category in the output layer. Training was done for 45 iterations, and the "best" network iteration ("*baseline*" network - ***B***) was determined by evaluation on the FIELD development-set. With this network a final test was also executed.

4.6 "Forced alignment"

Each waveform in the "baseline" hand-labeled training material plus the whole PANDA speech corpus (1041 digit sequences, 16247 digits) was then recognized using the best obtained network (B), with the result constrained to be the correct sequence of digits. This process, called "*forced alignment*", was used to generate time-aligned category labels. These force-aligned category labels were then used in a second cycle of training and evaluation was repeated to determine the new best network ("*force aligned*" network - ***FA***), which was finally evaluated with the same development and test data.

4.7 "Forward Backward" training

In order to explore the possibility to further improve the recognition results, the "*forward-backward*" (***FB***) training strategy was [15] recurrently applied (three times). 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) should be used. Then the forward-backward re-estimation algorithm could be used to regenerate the targets for the training utterances. The re-estimation can be 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.

4.8 Results

As illustrated in Table 3, various combination of features were considered, and, up to now, in terms of word and sentence recognition accuracy, the best obtained experimental results are those illustrated in Table 4 referring to 12 MFCCs with CMS, energy and corresponding delta values.

The global best network is that corresponding to the best network after the third Forward-Backward pass (FB3 – nnet17) characterized by a very high recognition accuracy, especially considering the high degradation level, in terms of background noise, channel noise or other non-speech phenomena, of the test-set speech material. In particular 99.72% WA and 99.15% SA was obtained on the development-set, and 98.68% WA and 90.76% SA on the test-set. Considering the test-set, these results correspond to 59% and 49% reduction in error compared to the performance obtained by IRST (96.8%) and CSELT (97.4%) respectively, at the word-level, and to 49% and 35% reduction in error compared with IRST (82.4%) and CSELT (85.7%) results at the sentence level [9-12].

		WA	SA
mfcc13(cms)+Δ	**dev**	**99.72**	**99.15**
	test	**98.68**	**90.76**
mfcc13(cms)+Δ+Δ²	**dev**	**99.82**	**99.15**
	test	**98.40**	**89.53**
mfcc7(cms)+Δ+Δ²	**dev**	**99.72**	**98.29**
	test	**97.88**	**86.24**
plp13(rasta)+mfcc13(cms)	**dev**	**99.54**	**96.58**
	test	**97.79**	**85.42**
plp13+mfcc13(cms)	**dev**	**99.63**	**97.44**
	test	**97.70**	**85.42**
plp13+mfcc13	**dev**	**99.54**	**96,58**
	test	**97.79**	**85.42**
plp9(rasta)+mfcc9(cms)	**dev**	**99.54**	**97.44**
	test	**98.01**	**87.06**
[plp9(rasta)+mfcc9(cms)]+Δ	**dev**	**99.72**	**98.29**
	test	**98.27**	**88.50**
[plp7(rasta)+mfcc7(cms)] +Δ	**dev**	**99.45**	**95.73**
	test	**98.01**	**87.68**

Table 3. Best recognition performances, in terms of "Word Accuracy" (WA) and "Sentence Accuracy" (SA) for various combination of features.

	HL (34)		**FA (21)**		**FB_1 (58)**		**FB_2 (19)**		**FB_3 (17)**	
	WA %	**SA %**	**WA %**	**SA %**	**WA %**	**SA %**	**WA %**	**SA %**	**WA %**	**SA %**
Dev	**99.72**	**98.29**	**99.72**	**99.15**	**99.54**	**97.44**	**99.54**	**97.44**	**99.72**	**99.15**
Test	**98.24**	**87.89**	**98.31**	**89.53**	**98.07**	**87.47**	**98.27**	**89.53**	**98.68**	**90.76**

Table 4. Recognition performance in terms of "Word Accuracy" (WA) and "Sentence Accuracy" (SA) for the best Hand-Labelled (HL - nnet-34), Forced-Alignment (FA - nnet-21) and Forward-Backward (FB1 - nnet-58, FB2 – nnet-19, FB3 – nnet-17). The best network for testing the system was choosen as the best FB3 network (nnet-17).

5. CONCLUSIONS

In summary, this work yielded a state-of-the-art telephone-channel Italian digit recognition system. The current-best Italian digit recognizer was implemented in the Toolkit's dialogue design module and a simple Italian-language demonstration program that accepts connected digit string or simple menu orders from a user has been created. These demonstration systems were installed on a laptop machine and were successful in informal presentations.

The present Italian “digits” recognizer will be included in the next version of the CSLU Speech Toolkit.

6. ACKNOWLEDGMENTS

The authors would like to sincerely thank IRST and CSELT companies for their cooperation in making available their corpora Field, Phone and Panda test-set. In particular, we would like to thank Gianni Lazzari, Daniele Falavigna, Roberto Gretter and Maurizio Omologo from IRST and Roberto Billi and Luciano Fissore from CSELT for their support and for their useful suggestions. Part of this work was made possible by the “International Short-Term Mobility Program” of Consiglio Nazionale delle Ricerche.

7. FUTURE RESEARCH

In preparation for future research, new software that allows a computer to record telephone-channel speech using the CSLU Toolkit to perform its basic functions was installed and tested with success. This software allows telephone-based interaction with the Toolkit as well as the collection of telephone-channel corpora for training new Italian recognition systems. Moreover a new package was developed and added to the Toolkit that will allow exploratory feature sets, which may currently require a great deal of computation time, to be easily integrated into the training and testing of an HMM/ANN recognizer. This will allow not only the development of full-scale recognition systems using these new features, but will also allow direct comparison of different feature sets given the same training procedures and corpora.

This work is inserted in a project whose aim is to contribute to the “Italianization” of the CSLU Toolkit and to support the dissemination of these tools and technologies.

8. REFERENCES

[1] Hermansky, H., “Perceptual Linear Predictive (PLP) Analysis of Speech,” Journal of the Acoustical Society of America, vol. 87, no. 4, pp. 1738-1752, April 1990.

[2] Davis, S. and Mermelstein, P., “Comparison of Parametric Representations for Monosyllabic Word Recognition,” IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-28, pp. 357-366, 1980.

[3] Hermansky, H., Morgan, N. 1994. RASTA Processing of Speech. IEEE Trans. on Speech and Audio Processing. Vol.2, No.4, 578-589.

[4] Furui, S. 1981. Cepstral Analysis Techniques for Automatic Speaker Verification. IEEE Transactions on Acoustic Speech and Signal Processing, Vol. 29, No. 2, 254-272.

[5] Fanty, M., Pochmara, J., and Cole, R.A. “An Interactive Environment for Speech Recognition Research”, ”, Proceedings of International Conference on Spoken Language Processing (ICSLP-92), Banff, Alberta, October 1992, 1543-1546.

[6] http://cslu.cse.ogi.edu/toolkit.

[7] Bourlard H., “Towards Increasing Speech Recognition Error Rates”. In *Proceedings EUROSPEECH95*, Madrid, Spain, September 1995, Vol. 2, pp. 883-894.

[8] P. Cosi and J.P. Hosom, “High Performance ‘General Purpose’ Phonetic Recognition for Italian”, Proceedings of International Conference on Spoken Language Processing (ICSLP-2000), Beijing, China, October 2000.

[9] B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter and M. Omologo, “Speaker Independent Continuous Speech Recognition Using an Acoustic-Phonetic Italian Corpus”, Proceedings of International Conference on Spoken Language Processing (ICSLP-94), Yokohama Japan, 1994.

[10] D. Falavigna and R. Gretter , “On Field Experiments of Continuous Digit Recognition over the Telephone Network”, Proceedings of EUROSPEECH '97, Rhodes Greece, 22-25 September 1997.

[11] Chesta, C., Laface, P. and Ravera, F. 1999. Connected Digit Recognition Using Short and Long Duration Models. In Proceedings of ICASSP-99, Phoenix, AZ, USA. March 15-19, 1999.

[12] M. Nigra, L. Fissore and F. Ravera, “Riconoscimento di Cifre Connesse su Rete Telefonica”, DT, Documenti Tecnici, CSELT.

[13] Fourcin A.J., Harland G., Barry W. and Hazan W., Eds. *Speech Input and Output Assessment, Multilingual Methods and Standards*, Ellis Horwood Books in Information Technology, 1989.

[14] W. Wei and S. Van Vuuren, “Improved Neural Network Training of Inter-Word Context Units for Connected Digit Recognition”. In Proceedings of International Conference on Acoustic Speech and Signal Processing (ICASSP ’98), Seattle, Washington, May 1998, Vol. 1, pp. 497-500.

[15] Yan, Y., Fanty, M. and Cole, R., “Speech Recognition Using Neural Networks with Forward-Backward Probability Generated Targets”. In *Proceedings ICASSP97*, April 1997, Vol. 4, pp. 3241-3244.