HIGH PERFORMANCE TELEPHONE BANDWIDTH SPEAKER INDEPENDENT CONTINUOUS DIGIT RECOGNITION

Piero Cosi, John-Paul Hosom**and Alberto Valente****

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

**Center for Spoken Language Understanding (CSLU-OGI)
Oregon Graduate Institute (OGI), P.O. Box 91000, Portland, Oregon 97291 USA (e-mail: 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 high-performance telephone-bandwidth speaker independent connected digit recognizer for Italian is described. The CSLU Speech Toolkit was used to develop and implement the hybrid ANN/HMM system, which is trained on context-dependent categories to account for coarticulatory variation. Various front-end processing and system architecture were compared and, when the best features (MFCC with CMS + Δ) and network (4-layer fully connected feed-forward network) were considered, there was a 98.92% word recognition accuracy and a 92.62% sentence recognition accuracy) on a test set of the FIELD continuous digits recognition task.

1. INTRODUCTION

The use of Automatic Speech Recognition (ASR) over standard and, more recently, cellular telephone lines has been steadily increasing over the past several years. For many applications of speech recognition over the telephone, such as credit card and account number validation, catalog ordering and digit dialing by voice, connected digit recognition (CDR) is absolutely essential.
Over the last two decades much progress has been made for recognizing spoken connected digit sequences recorded under very controlled non-telephone network condition [1-6]. CDR accuracies on the TI-database [7], as reported in the literature, have been superb with the best string accuracies almost approaching 100%. In spite of this, the recognition of connected digit strings spoken over telephone lines is a much more complicate task and, until now, it is not completely solved. The telephone network introduces so many effects on speech that the current ASR technology may not be so robust to let us achieve comparable recognition performance, even if continuous improvements have recently been obtained on English [8-10] and other languages such as Italian [11-13]. The "connected digit recognition" small-vocabulary task, with ten digits from "zero" through "nine", 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 [14] and MFC coefficients [15], together with two normalization procedures, such as RASTA [16] and Cepstral Mean Subtraction [17], were investigated.

2. RECOGNITION FRAMEWORK

The recognizer being described in this work was developed and implemented by the use of the CSLU Speech Toolkit [18] freely available through the CSLU OGI Web site [19]. The basic framework considered for recognition was that corresponding to an hybrid ANN/HMM architecture [20], 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 [12], and PANDA [11], [21]. 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 digit recognizer was trained on context-dependent categories to account for coarticulatory variations and recognizes any connected sequence of the 10 Italian digits (SAMPA transcription [22]):

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].

A 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/four-layer fully connected feed-forward network 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), 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. "silence" (.pau .garbage @br) and "closure" are 1-part units, "vowel" (i e E a O o u) is a 3-part unit, "unvoiced plosive " (t k) is 1-part right dependent unit, "voiced plosive" (d), "affricate" (dz tS), "fricative" (s v), "nasal" (n), "liquid retroflex"(r) and "glide" (w) are all 2-part units. 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 (see Table 1).

Group	Acoustic units in group	Description
$sil	.pau, .garbage @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 1. 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 [23]. 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 2^{nd} or 8^{th} percentile of all duration values. With a bigger percentile the mean duration for each phonetic category is increased thus reducing the probability to select short duration categories while increasing the probability of cancellation errors. 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 number of digits in any order, with optional silence between digits. In addition, a special word called "Garbage" was allowed at the beginning and end of each utterance to account for out-of-vocabulary sounds. This "garbage" word consisted of a single context-independent category (also called "garbage"); the value for this category was not an output of the neural network, but was computed as the Nth-highest output from the neural network at each frame [24]. In this study N was set to 5, if garbage model is included, and this value was empirically determined by varying N until a roughly equal error rate between insertions and deletions was obtained on our task. In our experiments N was also set to 100 setting in practice to 0 the probability of selecting a garbage category, simultaneously reducing cancellation errors.

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 [25] 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/four-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 2, 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 3 referring to 12 MFCCs with CMS, energy and corresponding delta values for a four-layer fully connected feed-forward network. After a forced alignment starting from the best network obtained in previous experiments [13] (see results on the left column of Table 2 for N=5 and 2nd percentile), the global best network is that corresponding to the best network after the second Forward-Backward pass (FB2 – nnet14) 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 97.44% SA was obtained on the development-set, and 98.92% WA and 92.62% SA on the test-set. Considering the test-set, these results correspond to 66% and 58% reduction in error compared to the performance obtained by IRST (96.8%) and CSELT (97.4%) respectively, at the word-level, and to 58% and 48% reduction in error compared with IRST (82.4%) and CSELT (85.7%) results at the sentence level [11][12][21].

		2nd percentile N=5		8th percentile N=100	
		WA	SA	WA	SA
mfcc13(cms)+ Δ	dev	99.72	99.15	99.72	97.44
	test	98.68	90.76	98.90	92.01
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	94.11	72.65		
	test	89.49	54.00		
plp13(cms)+ Δ	dev			99.54	97.44
	test			98.70	90.57
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		
[plp7(cms)+ mfcc7(cms)]+Δ	dev			99.82	98.29
	test			98.64	90.57
[plp13(cms)+ mfcc13(cms)]+Δ+Δ²	dev			99.63	98.29
	test			98.94	92.01
mfcc13(cms) +Δ	dev			99.72	97.44
	test			98.92	92.62

Table 2. Best recognition performances, in terms of "Word Accuracy" (WA) and "Sentence Accuracy" (SA) for various combination of features. Mean duration values are set as the 2nd [13] and 8th percentile and garbage is set as the Nth-highest output from the neural network at each frame. N=100 means in practice "no garbage model". The last raw refers to a four-layer fully connected feed-forward network (see details in Table 3).

	FA (14)		FB1 (14)		FB2 (14)		FB3 (29)	
	WA	SA	WA	SA	WA	SA	WA	SA
dev	99.82	98.29	99.82	98.29	99.72	97.44	99.82	98.29
test	98.68	90.98	98.72	92.01	98.92	92.62	98.81	91.19

Table 3. Recognition performance for a four-layer fully connected feed-forward network for the best Forced-Alignment (FA - nnet-14) and Forward-Backward (FB1 - nnet-14, FB2 – nnet-14, FB3 – nnet-29). The best network for testing the system was chosen as the best FB2 network (nnet-14).

5. CONCLUSIONS

In summary, this work yielded a state-of-the-art telephone-bandwidth Italian speaker independent connected digit recognition system. The current-best Italian digit recognizer was implemented in the CSLU Toolkit's dialogue design module and a simple real-time Italian-language demonstration program has been created. The present Italian digit 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.

7. REFERENCES

[1] G.R. Doddington, "Phonetically Sensitive Discriminants for Improved Speech Recognition", Proc. *IEEE-ICASSP*, 1989, pp.556-559.

[2] L.R. Rabiner, J.G. Wilpon, F.K. Soong, "High Prformance Connected Digit Recognition Using Hidden Markov Models", *IEEE Trans. ASSP,* Vol. 37, 1989, pp.1214-1225.

[3] J.G. Wilpon, C.H.Lee, L.R. Rabiner, "Improvements in Connected Digit Recognition Using Higer Order Spectral and Energy Features", Proc. *IEEE-ICASSP*, 1991, Vol. I, pp.349-352.

[4] J.L. Gauvain, C.H.Lee, "Improved Acoustic Modeling with Bayesian Learning", Proc. *IEEE-ICASSP*, 1992, Vol. I, pp.481-484.

[5] R. Haeb-Umbach, D. Geller, H. Ney, "Improvements in Connected Digit Recognition Using Linear Discriminant Analysis and Mixture Densities", Proc. *IEEE-ICASSP*, 1993, Vol. II, pp.239-242.

[6] Y. Normandin, R. Cardin, R. De Mori, "High Performance Connected Digit Recognition using Maximum Mutual Information Estimation", *IEEE Trans. SAP,* Vol. 2, N. 2, 1994, pp.299-311.

[7] L. Leonard, G.R. Doddington, "A Database for Speaker Independent Digit Recognition", Proc. *IEEE-ICASSP*, 1984 paper 42.11.

[8] E.R. Buhrke, R. Cardin, Y. Normandin, M,Rahim, J.G Wilpon, "Application of Vector Quantized Hidden Markov Modeling to Telephone Network Based Connected Digit Recognition", Proc. *IEEE-ICASSP*, 1994, Vol. I, pp. 105-108.

[9] D.L. Thomson, R. Chengalvarayan, "Use of Periodicity and Jitter as Speech Recognition Features", Proc. *IEEE-ICASSP*, 1998, Vol. I, pp. 21-24.

[10] J.P. Hosom, R.A. Cole, P. Cosi , "Evaluation and Integration of Neural-Network Training Techniques for Continuous Digit Recognition", Proc. *ICSLP-98*, 1998, Vol. 3, pp. 731-734.

[11] C. Chesta, P. Laface, F. Ravera, "Connected Digit Recognition Using Short and Long Duration Models", *Proc. IEEE-ICASSP*, 1999, pp.557-560.

[12] D. Falavigna, R. Gretter , "On Field Experiments of Continuous Digit Recognition over the Telephone Network", Proc. *EUROSPEECH*, 1997.

[13] P. Cosi, J.P. Hosom, F. Tesser, "High Performance Italian Continuous Digit recognition", Proc. *ICSLP-2000*, 2000, Vol. IV, pp. 242-245.

[14] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech", *JASA*, Vol. 87, N. 4, 1990, pp. 1738-1752.

[15] S. Davis, P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition", *IEEE Trans. ASSP*, Vol. 28, 1980, pp. 357-366.

[16] H. Hermansky, N. Morgan, "RASTA Processing of Speech", *IEEE Trans. SAP,* Vol. 2, No.4, 578-589.

[17] S. Furui, "Cepstral Analysis Techniques for Automatic Speaker Verification", I*EEE Trans. ASSP*, Vol. 29, No. 2, 1981, 254-272.

[18] M. Fanty, J. Pochmara, R.A. Cole, "An Interactive Environment for Speech Recognition Research", Proc. *ICSLP*, 1992, 1543-1546.

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

[20] H. Bourlard, "Towards Increasing Speech Recognition Error Rates", Proc. *EUROSPEECH*, 1995, Vol. 2, pp. 883-894.

[21] M. Nigra, L. Fissore, F. Ravera, "Riconoscimento di Cifre Connesse su Rete Telefonica", DT, Doc. Tecnici, CSELT.

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

[23] W. Wei, S. Van Vuuren, "Improved Neural Network Training of Inter-Word Context Units for Connected Digit Recognition", Proc. *IEEE-ICASSP,* 1998, Vol. 1, pp. 497-500.

[24] J.M. Boite, H. Bourlard, B. D'hoore, M. Haesen, "A New Approach Towards Keyword Spotting", Proc. *EUROSPEECH*, 1993, Vol. 2, pp. 1273-1276.

[25] Y. Yan, , M. Fanty, R. Cole, "Speech Recognition Using Neural Networks with Forward-Backward Probability Generated Targets", Proc. *IEEE-ICASSP*, 1997, Vol. 4, pp. 3241-3244.