Hybrid HMM-NN Architectures for Connected Digit Recognition

Piero Cosi

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

Abstract

This paper deals with the more recent results obtained by the application of the CSLU Toolkit frame-based hybrid HMM/ANN architecture on the connected digit recognition task for the Italian language. The hybrid architecture for speaker independent recognition is described and the last obtained results are introduced in detail.

1. Introduction

The "***digits***" small-vocabulary task is important for many telephone-based applications such as computer-assisted long-distance dialing or credit-card billing, requires extremely high accuracy, and focuses research on acoustic-level processing. Digits represent a tractable problem because the vocabulary is small and fixed, yet developing and optimizing performance on these recognizers is extremely important, since they are often used in spoken dialogue systems. In our previous work, high-performance recognition of English digits over the telephone channel and Italian digits over a microphone channel have been explored [1-2]. Various experiments have been carried out regarding the types of features that are used as input by the neural-network classifier, the types of context-dependent categories that are output by the classifier, and duration and grammar modeling [1]. The standard HMM and the hybrid HMM/ANN speech-recognition technology have been also compared, and it was found that the latest hybrid HMM/NN systems perform better, at least on this domain.

2. "Digits"

Three corpora have been used in this work in order to train, develop and test the telephone-channel digits recognition system: FIELD, PHONE [3] and PANDA [4]. 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, in fact, it corresponds to various credit-card-like digit strings pronounced by more than 1000 speakers. The speech material was divided in training, development and test sub-sets. Moreover, for a truly fair comparison with previous experimental results obtained by IRST [3], the test-set was chosen to be exactly the same speech material utilized in their experiments and it correspond to a subset (49%) of the FIELD corpus consisting of 488 digit sequences.

3. Hybrid HMM/NN

The platform for our work has been the CSLU Toolkit [5], which is freely available world-wide for research use[1], and includes software for signal processing, speech recognition, text-to-speech synthesis, facial animation, and dialogue design. The basic framework was a hybrid HMM/ANN architecture.

3.1 Feature Extraction

As illustrated in Figure 1 spectral features are computed at every 10-msec frame in the utterance and are input to a neural network for classification. The neural network receives not just the features for a given frame, but a set of

[1] The CSLU Toolkit is freely available for non-commercial use and may be downloaded from http://cslu.cse.ogi.edu/toolkit.

features for the given frame and a fixed, small number of surrounding frames. This "context window" of features is used to provide the network with information about the dynamics of the speech signal. In particular for the digit experiments the analysis vector was constructed with 12 MFCCs [6] and the energy parameter plus their delta values, and cepstral-mean subtraction (CMS) [7] was performed, with the mean computed using all frames of data. The context window comprises the features for frames at -60, -30, 0, 30 and 60 msec relative to the frame to be classified (for a total of 130 input values).

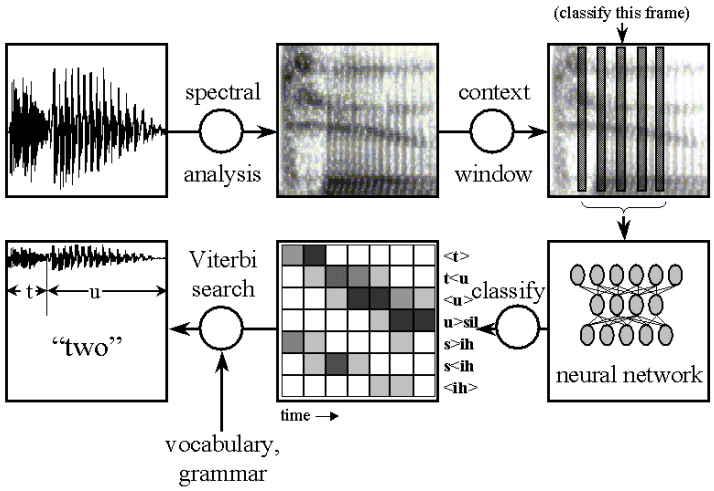

Figure 1. Graphical overview of the recognition process, illustrating recognition of the word "two".

3.2 Neural Network Architecture

At each frame, the neural network classifies the features in the context window into phonetic-based categories, estimating the probabilities of each category being represented by that set of features. The result of the neural network processing is a CxF matrix of probabilities, where C is the number of phonetic-based categories, and F is the number of frames in the utterance. In particular, in the digit case, the neural-network, that is simply 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.

3.3 Acoustic Units and 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, 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, .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 and 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.

The sequence of digits that best match this matrix of probabilities is determined using a Viterbi search, given the vocabulary and grammar constraints. In particular, a simple grammar [<any> (<digit> [silence])+ <any>] allowing any digit sequence in any order, with optional silence between digits, was considered. The search, as illustrated in Figure 2 for a simple two-word vocabulary, is usually thought of as traversing a state sequence, where each state represents a phonetic-based category, and there are certain probabilities of transitioning from one state to another.

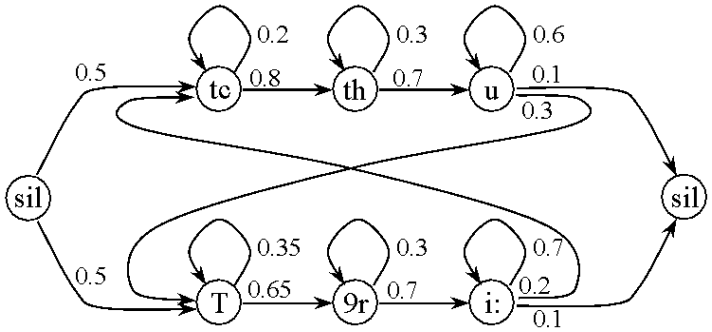

Figure 2. HMM state sequence for a two-word vocabulary.

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. Using a neural network to do this estimation has the advantage of not requiring assumptions about the distribution or independence of the input data, and neural networks easily perform discriminative training [8]. Also, neural networks can be used to perform recognition much faster than standard HMMs. A second difference is in the type of 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.

3.4 Neural Network Training

Neural-network training was done with standard back-propagation on a fully connected feed-forward network. The training was adjusted to use a negative penalty modification [9]. 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. 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. 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.

3.5 Training, Development and Test

In this work, the whole training procedure was done in three stages and, at each stage, evaluation was done on a development set of about 13% (127 digit sequences) of the FIELD corpus. At first training was done on the initially available hand-labeled phonetic transcriptions (***HL, Hand-Labelled training***), using binary target values for the neural network. Then on transcriptions that are automatically generated from the first stage using binary target values and the best *HL* network (***FA, Forced-Alignment training***). Finally, starting from the best *FA* network, the *forward-backward* re-estimation algorithm was used to regenerate the targets for the training utterances (***FB, Forward-Backward training***) [10]. As illustrated in Figure 3, like most of the other hybrid systems, the neural network is used as a state emission probability estimator. Unlike most of the existing hybrid systems, which do not explicitly train the within-phone relative likelihood, this new system trains the within-phone models to probability estimates obtained from the forward-backward algorithm, rather than binary targets. In other words this new training stage was executed using automatic transcriptions but with probabilistic target values obtained from the second stage. The re-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.

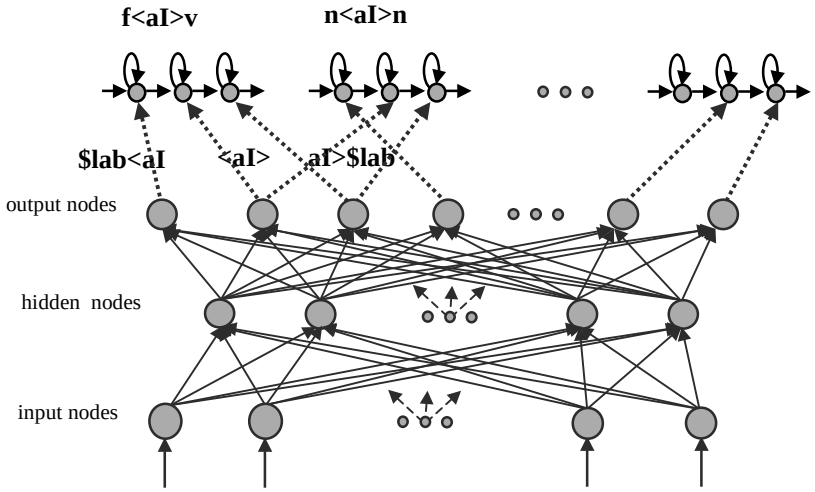

Figure 3. Overview of the hybrid system showing the relation between NN output nodes and the phone models.

4. Results

As illustrated in Figure 4, excluding those speech files included in the original FIELD test-set file list, ¾ of all remaining FIELD data, were used for *HL* training and the remaining ¼ was used for evaluation in the development stage. PHONE was entirely used for *HL* training too, while the whole PANDA corpus was added at the time in which *FA* training was considered.

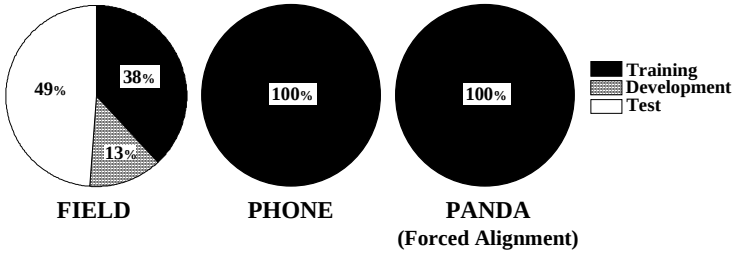

Figure 4. Training, Development and Test sets. As for FIELD, 38% of all the available data (388 *ds*, "*digit sequences*") was used for training, 13% (127 *ds*) was used for development and 49% (488 *ds*) was used for test. PHONE (2241 *ds* for a total of 9131 digits) and PANDA (1041 ds for a total of 16247 digits) were entirely utilized for training but the last one only for "Forced Alignment".

In this case results are illustrated in Table 3. 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 44% and 31% reduction in error compared with IRST (82.4%) and CSELT (85.7%) results at the sentence level [11-12].

		HL (34)		**FA (21)**		**FB1 (58)**		**FB2 (19)**		**FB3 (17)**	
		WA %	**SA %**	**WA %**	**SA %**	**WA %**	**SA %**	**WA %**	**SA %**	**WA %**	**SA %**
Dev	**FIELD**	**99.72**	**98.29**	**99.72**	**99.15**	**99.54**	**97.44**	**99.54**	**97.44**	**99.72**	**99.15**
Test	**FIELD**	**98.24**	**87.89**	**98.31**	**89.53**	**98.07**	**87.47**	**98.27**	**89.53**	**98.68**	**90.76**

Table 3. 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.

6. 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.

Acknowledgements

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.

References

[1] J. P. Hosom, R. A. Cole, and P. Cosi, "Improvements in Neural-Network Training and Search Techniques for Continuous Digit Recognition". *Australian Journal of Intelligent Information Processing Systems (AJIIPS)*, Vol. 5, N0. 4, Summer 1998, pp. 277-284.

[2] P. Cosi, and J. P. Hosom, "HMM/Neural Network-Based System for Italian Continuous Digit Recognition". *Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS '99)*, San Francisco, CA, USA, 14-18 August 1999. Vol. 3, pp. 1669-1672.

[3] 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, Vol. 4. pp. 1827-1830.

[4] C. Chesta , P. Laface and F. Ravera. "Connected Digit Recognition Using Short and Long Duration Models". *Proceedings of International Conference on Acoustic Speech and Signal Processing* (*ICASSP '99)*, Phoenix, AZ, USA. March 15-19, 1999.

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

[6] S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition". *IEEE Transactions on Acoustics, Speech and Signal Processing (ASSP)*, 1980, Vol. 28, pp. 357-366.

[7] S. Furui, "Cepstral Analysis Techniques for Automatic Speaker Verification". *IEEE Transactions on Acoustic Speech and Signal Processing (ASSP)*, Vol. 29, No. 2, 254-272.

[8] H. Bourlard, "Towards Increasing Speech Recognition Error Rates". *Proceedings of EUROSPEECH '95*, Madrid, Spain, September 1995, Vol. 2, pp. 883-894.

[9] 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.

[10] Y. Yan, M. Fanty and R.A. Cole, "Speech Recognition Using Neural Networks with Forward-Backward Probability Generated Targets". In *Proceedings of International Conference on Acoustic Speech and Signal Processing (ICASSP* '97), April 1997, Vol. 4, pp. 3241-3244.

[11] D. Falavigna and R. Gretter, "Riconoscimento di Cifre Connesse su Rete Telefonica", personal communication.

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