HMM/NEURAL NETWORK-BASED SYSTEM FOR ITALIAN CONTINUOUS DIGIT RECOGNITION

Piero Cosi* and John-Paul Hosom**

*Institute of Phonetics and Dialectology - C.N.R. Via G. Anghinoni, 10 - 35121 Padova (ITALY),
e-mail: cosi@csrf.pd.cnr.it www: http://www.csrf.pd.cnr.it

**Center for Spoken Language Understanding - Oregon Graduate Institute of Science and Technology
P.O. Box 91000, Portland Oregon 97291-1000 USA
e-mail: hosom@cse.ogi.edu www: http://cslu.cse.ogi.edu

ABSTRACT

An Italian speaker-independent continuous-speech digit recognizer is described. The CSLU Toolkit was used to develop and implement the system. In the first set of experiments, the SPK-IRST corpus, a collection of digit sentences recorded in a clean environment, was used both for training and testing the system. In the second set, a band-filtered version (between 300 Hz and 3400 Hz) of the SPK-IRST corpus was considered for training, while the telephone PANDA-CSELT corpus was used for testing the system. A hybrid HMM/NN architecture was applied; in this architecture, a three-layer neural network is used as a state emission probability estimator and the conventional forward-backward algorithm is applied for estimating continuous targets for the NN training patterns. The final network, trained to estimate the probability of 116 context-dependent phonetic categories at every 10-msec frame, was not trained on binary target values, but on the probabilities of each phonetic category belonging to each frame. Training and testing will be described in detail and recognition results will be illustrated.

1. INTRODUCTION

The recognition engine used in the experiments described in this paper is based entirely on the CSLU Toolkit[1], an integrated set of software and documentation that represents the state of the art in tools for research, development, and learning about spoken language systems [1]. In particular, it is based on a hybrid HMM/ANN framework [2, 3], in which a frame-based recognition strategy with context-dependent sub-phonetic states is adopted, where the state probability estimation is computed using a neural network.

2. CORPORA

The SPK-IRST [4] and the PANDA-CSELT [5] corpora are utilized in this work. The SPK-IRST is an Italian database of isolated and connected digits designed and collected at IRST (Istituto per la Ricerca Scientifica e Tecnologica, Trento, Italy), and conceived for speaker recognition and verification purposes [6]. The speech material considered for this work belongs to 40 speakers (19 females and 21 males) and most of the speakers are from the North-East of Italy. Twenty repetitions of the ten Italian digits and twenty different sequences of 8 randomly selected connected digits were recorded in a quiet room for each speaker during five recording sessions scheduled on different days[2]. Speech was acquired at 48 kHz with 16-bit accuracy, downsampled to 16 kHz and, finally, stored in SPHERE format[3] waveform files. Time-aligned phonetic transcriptions, labeled using the Speech Assessment Methods Phonetic Alphabet (SAMPA) [7], are provided for 10 speakers while word transcriptions are provided for each isolated and connected digit utterance in the corpus.

PANDA-CSELT was collected over the Italian Public Switched Telephone Network [5]. The data in this corpus were collected from thousands of people within various regions of Italy who recited their credit card number over the telephone in a natural speaking style. Because the data were collected from a large number of speakers from different backgrounds in different environments, the corpus contains a noticeable amount of aspects of "real-life" speech, including noise, widely-varying energy levels, dialect differences and other complications.

3. SYSTEM

The recognition system is based on the baseline CSLU-Toolkit frame-based approach illustrated in Figure 1. The waveform is divided into frames and specific features are computed for each frame. These features describe the spectral envelope of the speech at that frame and at a small number of surrounding frames. The features in each frame are classified into phonetic-based categories using a neural network. The outputs of the neural network are used as estimates of the probability, for each phonetic category, that the current frame contains that category. The matrix of probabilities and a set of pronunciation models is used by a Viterbi search to determine the most likely words.

The neural network is trained to estimate the posterior probabilities of context-dependent phonetic categories which, given a certain lexicon, can be determined from the phonetic-level pronunciation models, the groupings of phones into clusters of similar phones, and the number of parts to split each phoneme into. One-part phonemes are context independent. Two-part phonemes have the left (first) half dependent on the preceding phoneme and the right (last) half dependent on the following phoneme. Three-part phonemes have a left third that is dependent on the preceding phoneme, a middle third that is context

independent, and a right third that is dependent on the following phoneme.

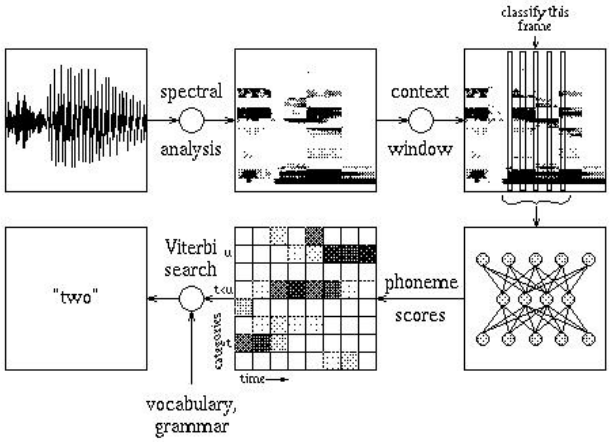

Figure 1. Overview of baseline CSLU-Toolkit frame-based speech recognition using neural networks.

4. EXPERIMENTS

In the first experiment (**exp1**), the clean SPK-IRST corpus (**C1**) was used for training, developing and testing the system. Three-fifths of the available data were randomly chosen and allocated (in a speaker-pooled way) for training (**C1-train**), and one-fifth each was allocated for development (**C1-dev**) and testing (**C1-test**). In the second experiment (**exp2**), a band-filtered version (between 300 Hz and 3400 Hz) of the SPK-IRST corpus (**C2**) was considered for training and development. The 10 speakers for which hand labels were available were used for training (**C2-train**), while the remaining 30 speakers were used for developing (**C2-dev**). A subset of the telephone PANDA-CSELT corpus (**C3**) [5] was used for testing the system (**C3-test**). It is quite evident that recognition performance in this case will greatly suffer the critical mismatch between training and testing speech material. The Italian digit lexicon and grammar to be recognized in both experiments are illustrated in Table 1.

word	pronunciation
zero	{dz E r o}
uno	{u n o}
due	{d u e}
tre	{t r E}
quattro	{k w a tt r o}
cinque	{tS i n k w e}
sei	{s E I}
sette	{s E tt e}
otto	{O tt o}
nove	{n O v e}
separator	{.pau [.garbage] .pau}

$digit	zero | uno | due | tre | quattro | cinque | sei | sette | otto | nove
$grammar	[separator%%] < $digit [separator%%] > [separator%%]

Table 1. Italian digit lexicon and grammar.

4.1 Segmentation

A three-layer neural network was trained to estimate, at every 10-msec frame, the probability of 116 context-dependent phonetic categories. These categories are created by splitting each phoneme, as illustrated in Table 2, into one, two, or three parts, depending on the length of the phoneme and how much the phoneme was thought to be influenced by coarticulatory effects. Phoneme states were trained for different preceding and following phonetic contexts, and some phonetic contexts were grouped together to form a broad-context grouping. For example, the left part of fricative /s/ and affricates /tS/, and /dz/ were combined, as illustrated in Table 3, into one broad-context category. The broad-context groupings were done based on acoustic-phonetic knowledge.

phone	parts
.pau	1
n	2
r	2
s	2
v	2
w	2
d	2
t	2
k	2
tt	2

phone	parts
tS	2
dz	2

u	3
o	3
O	3
a	3
E	3
e	3
I	3

Table 2. Phones and number of parts to split each phone into, for the Italian digit lexicon.

group	phones in group	description
$sil	.pau, .garbage	silence
$udp_l	t, tt	unvoiced burst to the left
$udp_r	t, tt, tS	unvoiced closure to the right
$vdp_l	d	voiced burst to the left
$vdp_r	d, dz	voiced closure to the right
f_l	s, tS, dz	frication to the left
f_r	s	frication to the right
$bck	u, o, O	back vowels
$mid	a, E	mid vowels
$frn	i, e	front vowels

Table 3. Groupings of phones into clusters of similar phones.

4.2 Feature Extraction

A combination of 13 Perceptual Linear Predictive Coefficients (PLPCs) [8] and 13 Mel Frequency Cepstral Coefficients (MFCCs) [9] were computed using RelAtive SpecTrAl (RASTA) [10] analysis and Cepstral Mean Subtraction (CMS) [11] pre-processing techniques, respectively. The combination of PLP and MFCC features was motivated by the hypothesis that training with the two slightly different representations would provide somewhat more robustness to noise, and that the combination of RASTA (which emphasizes regions of transition) and CMS (which does not emphasize transitions) would provide complimentary information.

4.3 Baseline Training

At each frame, a 130 dimensional vector of PLPCs+MFCCs was

constructed using five surrounding frames; 13 PLPCs and 13 MFCCs from frames at -60, -30, 0, 30, and 60 msec relative to the frame of interest were considered.

The training data were searched to find all the vectors of each category in the hand-labeled section of *C1* in *exp1* and of *C2* in *exp2*. The neural network was trained using the back-propagation method with 130 inputs, 200 nodes in the single hidden layer, and one node for each context-dependent category in the output layer (for a total of 116 output nodes). Training was done for 30 iterations, and the iteration with the best performance on the development set was chosen to be the final baseline neural network. This network, which will be referred to as *'baseline*' network **B** (**B1** or **B2** in the case of experiment 1 or 2 respectively), was finally evaluated with the test speech material (*C1-test* and *C3-test* for experiments 1 and 2, respectively).

4.4 Forced alignment

The SPK-IRST corpus we want to train on has been completely orthographically transcribed but only partially phonetically segmented and labeled. In this case, we can create either phonetic labels or category labels for the entire training material using a process called "forced alignment". Forced alignment is the process of using an existing recognizer to recognize a training utterance, where the grammar and vocabulary are restricted to be the correct result. The result of forced alignment is a set of time-aligned labels that give the existing recognizer's best alignment of the correct phonemes or categories. If the existing recognizer is good, then the labels will have more consistent time alignments than the hand labels. These labels can then be used for training a new recognizer. Even if the existing recognizer produces some alignment errors, this process can be used to determine an initial set of labeled training data. The best previously-trained networks *B1* and *B2* are, in fact, utilized to force align all of the training data in *C1* and *C2*[4], respectively, and a new "force-aligned" network is trained for a certain number of iterations, ranging from 30 to 60, using all of these new phonetically aligned data. The best force-aligned network, as evaluated on the development set, is chosen to be the final force-aligned neural network, called **FA** (**FA1** or **FA2** in the case of experiment 1 or 2 respectively) and is finally evaluated with the testing speech material (*C1-test* and *C3-test* for experiments 1 and 2, respectively).

4.5 Recognition

For recognition of an utterance, PLPCs+MFCCs vectors are computed in the same way as for training. These vectors are input to the neural network, which computes for each frame the probabilities that the current frame contains each of the specified categories. As illustrated in Figure 1, the result of classification is therefore a C x F matrix of probabilities, where C is the number of categories and F is the total number of frames. This matrix is then used by a Viterbi search algorithm to determine the most likely sequence of words. The Viterbi search uses minimum and maximum durations of each category to constrain the possible word choices, but these are not "hard" limits. If the duration of a hypothesized category falls beyond one of the specified limits, a penalty is applied; this penalty is proportional to the time difference between the specified limit and the hypothesized duration. Initial values for these limits are taken from the durations of the categories that were used to train the baseline network. These values are refined during the development stage by taking durations of the categories that were created during forced alignment.

4.6 Testing

The described recognition strategy is applied in *exp1* with the subset of *C1* allocated for testing, while in experiment 2 the testing material *C3* refers to the true telephone PANDA-CSELT corpus.

4.7 Forward Backward training

For *exp2*, the '*forward-backward*' (***fb***) training strategy [2] was applied in order to explore the possibility to further improve the recognition results. 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 was used, 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. This new configuration was called **Forward-Backward** (**FB**) neural networks. To start FB training for *exp2* (**FB2**), an initial binary-target neural network is required. For this initial network, we used the network resulting from forced-alignment training (*FA2*). Then the *forward-backward* re-estimation algorithm is used to regenerate the targets for the training utterances. The re-estimation is 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 are trained using the standard stochastic back-propagation algorithm, with mean-square-error as the cost function.

5. RESULTS

5.1 Experiment 1

In the first experiment the *baseline* network is found using speech material sets belonging to the clean SPK-IRST (*C1*) corpus as indicated in Table 4. Corresponding recognition results are shown in Table 5.

training	**hand-labeled section of *C1-train* (3/5 of SPK-IRST clean)**
development	**1/5 of *C1-train***
testing	**1/5 of *C1-train***

Table 4. Training, development and testing sets for *baseline* network in *exp1*.

B1	**Itr #**	**Snts #**	**Wrds #**	**Sub %**	**Ins %**	**Del %**	**WrdAcc %**	**SntAcc %**
dev	**20**	**1720**	**3120**	**0.22**	**0.00**	**0.16**	**99.62**	**99.30**
test	**20**	**1720**	**3120**	**0.13**	**0.06**	**0.16**	**99.65**	**99.42**

Table 5. *Baseline* recognition results for *exp1*.

Training, development, and testing sets referring to the *force aligned* network for *exp1* are illustrated in Table 6, while corresponding recognition results are shown in Table 7.

training	*C1-train* (force aligned with B1)
development	1/5 of *C1-train*
testing	1/5 of *C1-train*

Table 6. Training, development and testing sets for *force-aligned* network in *exp1*.

FA1	Itr #	Snts #	Wrds #	Sub %	Ins %	Del %	WrdAcc %	SntAcc %
dev	38	1720	3120	0.16	0.00	0.03	99.81	99.65
test	38	1720	3120	0.16	0.10	0.10	99.65	99.53

Table 7. *Force-aligned* recognition results for *exp1*.

5.2 Experiment 2

In the second experiment the *baseline* network is found using speech material sets belonging to the band-filtered version of SPK-IRST (*C2*) as indicated in Table 8. Corresponding recognition results are shown in Table 9.

training	hand-labeled section of *C2* 10 speakers, 1/4 of band-filtered SPK-IRST
development	30 speakers, 3/4 of *C2*
testing	*C3* (PANDA-CSELT)

Table 8. Training, development and testing sets for *baseline* network in *exp2*.

B2	Itr #	Snts #	Wrds #	Sub %	Ins %	Del %	WrdAcc %	SntAcc %
dev	29	4950	8800	0.26	0.10	0.10	99.53	99.29
test	29	990	15483	3.44	4.04	0.67	91.86	51.11

Table 9. *Baseline* recognition results for *exp2*.

Training and development sets referring to the *force aligned* and to the *forward-backward* networks for *exp2* are illustrated in Table 10, while corresponding recognition results are shown in Table 11.

training	*C2* (force aligned with *B2*)[4]
development	*C3* (PANDA-CSELT)

Table 10. Training and development sets for *force-aligned* and *forward-backward* networks in *exp2*.

FA2	Itr #	Snts #	Wrds #	Sub %	Ins %	Del %	WrdAcc %	SntAcc %
dev	22	990	15483	2.98	4.31	0.50	92.21	55.15

FB2	Itr #	Snts #	Wrds #	Sub %	Ins %	Del %	WrdAcc %	SntAcc %
dev	21	990	15483	2.42	4.56	0.47	92.55	53.74

Table 11. *Force-aligned* and *forward-backward* recognition results for *exp2*.

6. DISCUSSION

Very good results have been obtained for a speaker-independent continuous digit recognition task (*exp1*) in a clean environment. Encouraging results have been achieved on a similar task using the same clean corpus for training but a much more difficult telephone-band environment for testing (*exp2*), suggesting the effectiveness of the CSLU Toolkit in building real-life speech recognition systems. Currently we are extending this work by adopting two true telephone-band corpora for training the system, in order to cover more 'real life' complications, such as those encountered in the testing *C3* corpus, and we are very confident that results will be highly improved. Once the development set results have reached an acceptable level, we will perform final test-set results with the FB network. To compensate for the noise effects, methods based on spectral subtraction can also be applied [12]. This consists of subtracting a spectral estimate of the noise from each short time speech spectrum, but, due to the noise variability across different telephone calls, the noise estimate has to be updated during each call.

ACKNOWLEDGMENTS

This work has been made possible thanks to various support from people of CSLU. Among them we would like to thank Ronald Cole, Hynek Hermansky, Johan Shalkwyk, Stephen Sutton and Jacques de Villieres. John-Paul Hosom was supported by an NSF GRT grant, GER-9354959.

NOTES

1. The CSLU Toolkit is freely available for non-commercial use and may be downloaded from http://cslu.cse.ogi.edu/toolkit.
2. The isolated part of the whole SPK corpus, containing isolated digits collected from 100 speakers, is released on a CD-ROM by ELRA [4].
3. The SPHERE software package is public domain and the source code is available by anonymous ftp from: ftp://jaguar.ncsl.nist.gov/pub/sphere_2.6a.
4. In the case of experiment 2, the force-alignment is executed on the whole C2 (SPK-IRST band-filtered).

REFERENCES

[1] Fanty, M., Pochmara, J., and Cole, R.A. 1992. An Interactive Environment for Speech Recognition Research. In *Proceedings of ICSLP-92*, Banff, Alberta, October 1992, 1543-1546.

[2] Yan, Y., Fanty, M., and Cole, R.A. 1997. Speech Recognition Using Neural Networks with Forward-Backward Probability Generated Targets. In *Proceedings of ICASSP-97*, Munich, Germany, April 21-24, 1997, 3241-3244.

[3] Cosi, P., Hosom, J.P., Shalkwyk, J., Sutton, S., and Cole, R.A. 1998. Connected Digit Recognition Experiments with the OGI Toolkit's Neural Network and HMM-Based Recognizers. In *Proceedings of IEEE IVTTA-ETWR-98*, Turin, Italy, September 29-30, 1998, pp. 135-140.

[4] From the World Wide Web. 1998. European Language Resources Association: *http://www.icp.grenet.fr/ELRA/cata/spee_det.html#spk*.

[5] 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 (to be published).

[6] Brunelli, R. and Falavigna, D. 1995. Person Recognition Using Multiple Cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, Vol. 17, No. 10, 955-966.

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

[8] Hermansky, H. 1990. Perceptual Linear Predictive (PLP) Analysis of Speech. *Journal of Acoustical Society of America*, Vol. 87, No. 4, 1738-1752.

[9] Davis, S.B. and Mermelstein, P. 1990. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics Speech and Signal Processing*. Vol. 28, No. 4, 357-366.

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

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

[12] Acero, A. 1993. Acoustical and Environmental Robustness in Automatic Speech Recognition. Kluwer Academic Publishers, Boston.