Towards the Italian CSLU Toolkit

Piero Cosi[*], *John-Paul Hosom and Fabio Tesser*[***]

[*]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

[**]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/

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

Abstract

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.

On the other hand, in many other tasks a speaker-independent domain-specific vocabulary (such as "collect call", "calling card", "operator", or "help") needs to be recognized. For such tasks, a "***general-purpose***" (***gp***) recognizer that is capable of recognizing all permissible phoneme strings in a language is required.

The more recent results obtained by the application of the CSLU Toolkit frame-based hybrid HMM/ANN architecture on these recognition tasks for the Italian language are described. 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.

1. Introduction

In our previous work, high-performance recognition of English digits over the telephone channel and Italian digits over a microphone channel have been explored [1-4]. 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 [4]. The standard HMM speech-recognition technology and the hybrid HMM/ANN systems in use at CSLU have been also compared, and it was found that the latest hybrid NN/HMM systems perform better, at least on this domain.

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. Moreover, while the tasks are tractable, they present significant research challenges. On the other hand, a general-purpose (GP) speaker- and vocabulary-independent recognizer is necessary for rapid prototyping of spoken language systems for arbitrary tasks. Moreover the GP recognizer enables recognition of arbitrary words or phrases.

2. CSLU Toolkit

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 for the Toolkit's hybrid HMM/ANN speech recognition systems is illustrated in Figures 1 and 2. The system uses features that represent the spectral envelope (warped to emphasize the perceptually-relevant aspects [6, 7]) and its energy given a fixed window size. These spectral features are computed at every 10-msec frame in the utterance and are input to the neural network for classification. The neural network receives not just the features for a given frame, but a set of 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. 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.

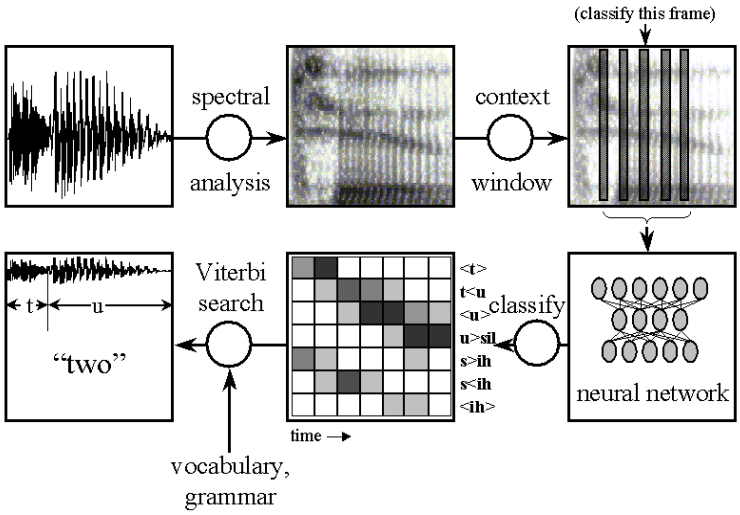

Figure 1. Graphical overview of the recognition process, illustrating recognition of the word "two".

The word or words that best match this matrix of probabilities is determined using a Viterbi search, given the vocabulary and grammar constraints. The search is usually thought of as traversing a state sequence (illustrated in Figure 2 with a simple two-word vocabulary), where each state represents a phonetic-based category, and there are certain probabilities of transitioning from one state to another.

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

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

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.

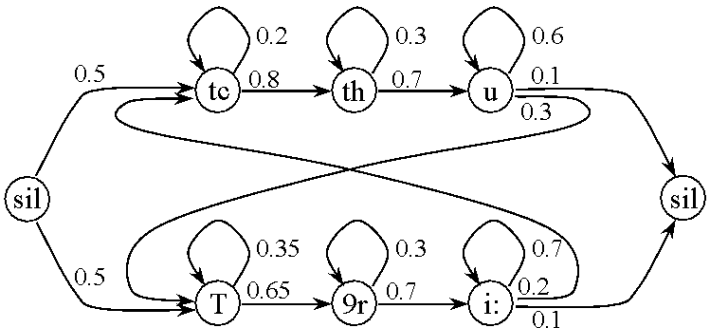

Figure 2. HMM state sequence for a two-word vocabulary.

3. Experiments

The present work is focused on the development of a digit-recognition system for Italian over telephone channels and of an Italian general-purpose recognizer for microphone channels.

3.1 "Digits"

3.1.1 Corpus

Two corpora have been used for training and development of the telephone-channel digits recognition system: the FIELD corpus and the PHONE corpus [9], both graciously provided by IRST as part of a research agreement. Both corpora had been transcribed at the word and phoneme level, with the time locations of each word and phoneme determined by an automatic procedure and then manually adjusted. Our intention is to create a digit recognizer that has been trained on these two corpora, and then perform final test evaluation on these corpora as well as the CSELT PANDA corpus [10]. In particular for the three corpora, Training, Development and Test sections were organized as illustrated in Figure 3.

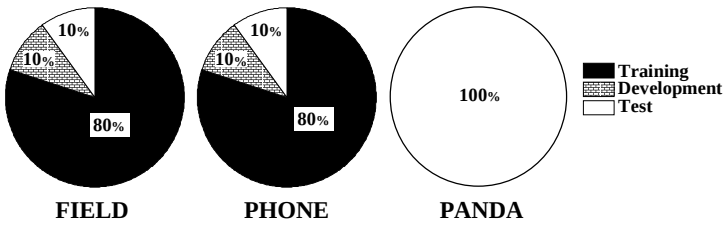

Figure 3. Training, Development and Test sets for FIELD (*F*), PHONE (*P*) and PANDA (*PA*). In particular, 721 digit sequences (*ds*) (6790 digits) in *F* and 1842 *ds* (7504 digits) in *P* were used for training. 85 *ds* (791 digits) in *F* and 191 *ds* (791digits) in *P* were used for development. 88 *ds* (809 digits) in *F*, 208 *ds* (836 digits) in *P* and 1041 *ds* (16247 digits) in *PA* were used for test.

If acceptable performance can be obtained on the PANDA corpus, then this indicates that the recognizer has successfully learned the digits task without being "tuned" to the corpora used in training. 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. The PHONE corpus has a large number of hesitations, breath noise, and

other "spontaneous speech phenomena", and has been divided into high-quality, medium-quality, and low-quality sections, depending on the degree of such phenomena in each utterance [9]. As the low-quality section contains mostly out-of-vocabulary words, and as our evaluation was restricted to in-vocabulary words, we did not evaluate on the low-quality section of the PHONE corpus.

3.1.2 Feature Extraction

As for feature extraction, 13 MFCC [7] features (12 cepstral coefficients and 1 energy parameter) plus their delta values are continuously computed with a 10-msec frame rate, as illustrated in the overview of the full procedure in Figure 4. Cepstral-mean subtraction (CMS) [11] was performed, with the mean computed using all frames of data.

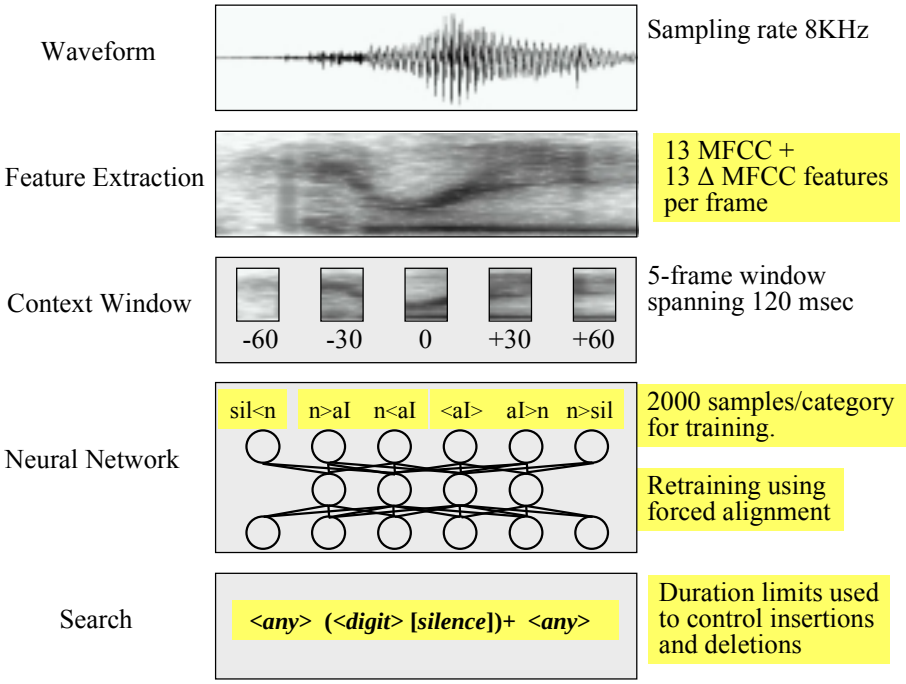

Figure 4. Overview of the full procedure.

3.1.3 Neural Network Architecture

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 (for a total of 130 input values) (see Figure 4). The neural-network is simply a three-layer fully connected feed-forward network.

3.1.4 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 [12]. 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.

3.1.5 Acoustic Units and Categories

A three-layer neural network, with 130 inputs and 200 nodes in the single hidden layer, was trained to estimate, at every 10-msec frame, the probability of 116 context-dependent phonetic categories. These categories are 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.

A simple grammar [<any> (<digit> [silence])+ <any>] allowing any digit sequence in any order, with optional silence between digits (see Figure 4), was considered.

3.1.6 Training, Evaluation and Test

In this work, training was done in three stages and, at each stage, evaluation was done on a development set of about 800 digits from each corpus.. At first training was

done on the initial 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***) [13]. As illustrated in Figure 5, like most of the other hybrid systems, the neural network is used as a state emission probability estimator and, unlike most of the existing hybrid systems, which do not explicitly train the within-phone relative likelihoods, 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.

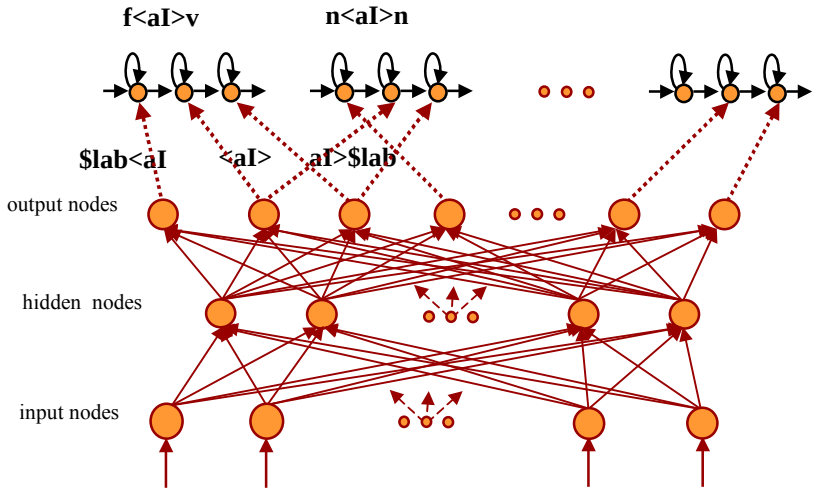

Figure 5. Overview of the hybrid system showing the relation between NN output nodes and the phone models.

3.1.7 Results

The system was evaluated with the FIELD and PHONE development set and tested with the FIELD, PHONE and PANDA test set, and results are illustrated in Table 3.

		HL (28)		FA (42)		FB (21)	
		WA %	SA %	WA %	SA %	WA %	SA %
Dev	FIELD	99.37	95.29	99.37	95.29	99.49	96.47
	PHONE	97.09	91.62	97.72	93.19	97.22	92.15
Test	FIELD			99.75	97.73		
	PHONE			98.68	95.19		
	PANDA			98.60	84.82		

Table 3. Recognition performance in terms of "Word Accuracy" (WA) and "Sentence Accuracy" (SA) for the best *Hand-Labelled* (HL) network-28, *Forced-Alignment* (FA) network-42 and *Forward-Backward* (FB) network-21. The best network for testing the system was choosen as the best FA network (nnet-42) given that FB performance were slightly worse.

The best network was chosen as the 42nd network after FA training given that it gave slightly better results (considering both FIELD and PHONE evaluation sets)

comparing to those obtained after FB training. A *word-level accuracy* (WA) of 99.75% and a *sentence-level accuracy* (SA) of 97.73% were achieved on the FIELD test set, while a word-level accuracy of 98.68% and a sentence-level accuracy of 95.19% were achieved on the PHONE test set. These results are quite better than those on the CSLU 30Knumbers telephone-channel continuous English digits task, with best known performance of around 98%, and they represent the best results obtained so far on these data, as reported by IRST [14] and CSELT [15]. In fact, at the word-level, they correspond to 92% and 71% reduction in error compared to the performance obtained by IRST on the FIELD (96.8) and PHONE (95.5%) corpus respectively [14]. Considering CSELT's performance, they are instead 90% and 72% reduction in error relatively to the FIELD (97.4%) and PHONE (95.2%) corpus respectively [15]. The final test on PANDA test set, with the same best FA network resulted in word-level accuracy of 98.6% and in sentence-level accuracy of 84.82. At the word level, this represents 53% reduction in error compared to the performance obtained by IRST on PANDA (97.0%) and 55% increase in error considering CSELT's best performance (99.1%) on the same corpus [10], [15]. However in CSELT's experiment training material was quite incomparable, in terms of quantity (8539 "*credit card*" digit sequences from the same PANDA corpus for training [10]), and in term of quality (training material belongs to the same "credit card" domain of the test material) with that available for this work.

3.2 "FIELD Digits"

Finally, we recently received by IRST more data belonging to the FIELD corpus and also the original FIELD test-set file list utilized in their experiments whose results are reported in [14]. Thus, for a truly fair comparison, a new experiment was organized in order to test the system with these new FIELD data. In this case, as illustrated in Figure 6, 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. The results obtained in this experiment are illustrated in Table 4.

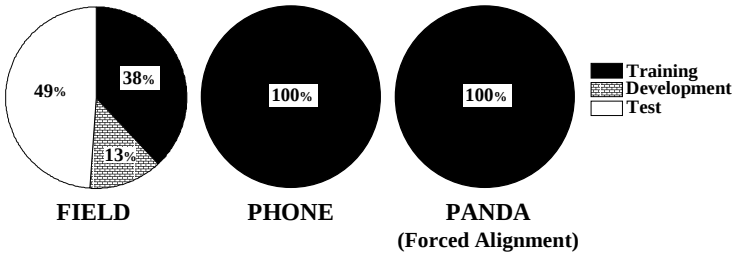

Figure 6. Training, Development and Test sets for the "*FIELD digits*" experiment. As for FIELD, 38% of all the available data (383 digit sequences) was used for training, 13% (127 digit sequences) was used for development and 49% (488 digit sequences) was used for test. PHONE and PANDA were entirely utilized for training but the last one only for "Forced Alignment".

In this case results are worse than those obtained in the previous experiment (see Table 3. WA 99.75%, SA 99.73%). This is probably due the fact that the new FIELD test-set material (488 digit sequences) is bigger than that utilized in the "digits" case (88 digit sequences, see Figure 3). Moreover these new data seems also more

degraded in terms of background noise, channel noise or other non-speech phenomena. However these values still correspond, at the word-level, to 47% and 35% reduction in error compared to the performance obtained by IRST (96.8) and CSELT (97.4%) respectively.

		HL (34)		FA (21)		FB (58)	
		WA %	SA %	WA %	SA %	WA %	SA %
Dev	FIELD	99.72	98.29	**99.72**	**99.15**	99.54	97.44
Test	FIELD	98.24	87.89	**98.31**	**89.53**	98.07	87.47

Table 4. Recognition performance in terms of "Word Accuracy" (WA) and "Sentence Accuracy" (SA) for the best *Hand-Labelled* (HL) network-34, *Forced-Alignment* (FA) network-21 and *Forward-Backward* (FB) network-58. The best network for testing the system was choosen as the best FA network (nnet-21) given that FB performance were slightly lower then FA performance.

3.3 "General Purpose"

Although the *digits* task is an important one, in many tasks a speaker-independent domain-specific vocabulary (such as "collect call", "calling card", "operator", or "help") needs to be recognized. For such tasks, a general-purpose recognizer that is capable of recognizing all permissible phoneme strings in a language is required.

3.3.1 Corpus

To train, develop and test such a recognizer, the APASCI [16] corpus from ELRA [17] has been considered. This corpus contains nearly 4000 sentences read by over 150 speakers, where the sentences have been designed to maximize the number of phonemes occurring in different contexts. The ELRA-provided corpus comes with a set of transcriptions at the word and phoneme level thus a recognizer was trained on the APASCI corpus with these transcriptions. In particular, 1250 hand-labeled sentences were used for training, 105 were considered for the development stage and 715 for the test phase, as illustrated in Figure 7.

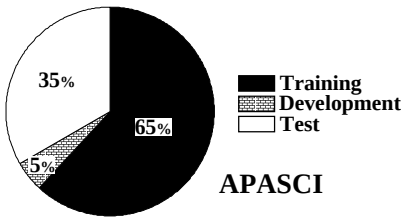

Figure 7. Training, Development and Test sets for APASCI In particular, 1250 sentences were used for training, 105 sentences for development and 715 sentences for test.

3.3.2 Feature Extraction

3.3.3 Neural Network Architecture

3.3.4 Neural Network Training

As for Feature Extraction, Neural Network Architecture and Neural Network Training, the considerations of section 3.1.2, 3.1.3, 3.1.4, for the *digits* case, apply in the *general purpose* case.

3.3.5 Acoustic Units and Categories

In perfect analogy with the *digits* case, a three-layer neural network, with 130 inputs and 250 nodes in the single hidden layer, was trained to estimate, at every 10-msec frame, the probability of 545 context-dependent phonetic categories. These categories are created by splitting each acoustic unit (*AU*), as illustrated in Table 5 and 6, 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.

Acoustic units	Parts	Description
.pau	1	silence
i e E a O o u	3	unstressed vowel
ii ee EE aa OO oo uu	3	stressed vowel
pcl bcl tcl dcl kcl gcl	1	closure
p b t d k g	r	plosive
ts dz dZ tS	2	affricate
s z f v S	2	fricative
m n N	2	nasal
l r L	2	liquid
j w	2	glide
@sch	2	schwa

Table 5. Acoustic units (SAMPA, except closures) and number of parts to split each unit into, for the Italian "general purpose" recognizer (r means "*right dependent unit*").

Similarly to the *digits* case, *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.

Group	Acoustic units in group	Description
$sil	.pau .garbage	silence
$fnt	i ii e ee j	front
$mid	E EE a aa @sch	mid
$bck	O OO o oo u uu w	back
$lab	p b f v m pcl bcl	labial
$alv	t d ts dz s z n tcl dcl	alveolar
$pal	dZ tS S N L	palatal
$vel	k g kcl gcl	velar
$lat	l	lateral
$ret	r	retroflex

Table 6. Groupings of acoustic units into clusters of similar units, for the Italian "general purpose" task.

3.3.6 Training, Evaluation and Test

The training data were searched to find all the vectors of each category in the hand-labeled training section of APASCI and evaluation was done on the corresponding development set. Training was done following the same scheme utilized in the *digits* case in which *Hand-Labelled training* is followed by *Forced-Alignment* and F*orward-Backward training*.

3.3.7 Results

As of the time of this writing, two of the three stages have been completed: *HL-* and *FA-training*. As illustrated in Table 7, phoneme-level accuracy of 82.90 and 80.53% on the APASCI development and test set respectively has been obtained.

	Itr #	Snts #	Wrds #	Sub %	Ins %	Del %	PhnAcc %
dev	**24**	**105**	**5235**	**10.41**	**2.56**	**4.45**	**82.90**
test	**24**	**715**	**36439**	**11.97**	**3.24**	**5.12**	**80.53**

Table 7. Recognition performance in terms of phone accuracy for the development and test set.

This level of accuracy is much greater than on a similar English-language corpus (with state-of-the-art performance of slightly better than 70%) and it represents the best performance obtained so far on this corpus, with no grammar and no phonotactic constraints. In fact, the performance obtained so far by IRST on an extended version of the same APASCI corpus [16] range, at the phone level, from 71.34% to 79.04%, while considering context-independent units (CIUs) and from 75.38 to 76.60 with Syllable-type units (SUs). When context-dependent units (CDUs) were considered, results were slightly better than ours, ranging from 81.36 to 82.44. However in this case, in contrast with our present implementation, phonotactic constraints were introduced in order to inhibit the recognition of unit sequences having incompatible contexts and this, according to the authors, improved accuracy from 2% to 3% depending on the particular unit set. Moreover in this case a very complex and sophisticated HMM system, with 16 gaussian mixtures per state and a large number (from 337 to 849) of context-dependent states was used in comparison to the rather straightforward architecture of the system being described in this work.

4. Conclusions

In summary, this work yielded a state-of-the-art telephone-channel Italian digit recognition system and excellent performance on microphone-channel Italian general-purpose recognition and further development of these systems will hopefully improve results even more. The results obtained on FIELD and PHONE corpora represent the best recognition performance obtained so far on these data. On CSELT PANDA corpus results were comparable, but slightly worse, with those obtained by CSELT. However, CSELT training material was quite incomparable, in terms of quantity (8539 "*credit card*" digit sequences from the same PANDA corpus for training), and in term of quality (training material belongs to the same "credit card" domain of the test material) with that available for this work.

The current-best Italian digit and general purpose recognizers were 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.

5. 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.

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] P. Cosi, J. P. Hosom, J. Shalkwyk, S. Sutton, and R. A. Cole, "Connected Digit Recognition Experiments with the OGI Toolkit's Neural Network and HMM-Based Recognizers". *Proceedings 4th IEEE Workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA-ETRW '98)*, Turin, Italy, 29-30 September 1998, pp. 135-140.

[2] J. P. Hosom, P. Cosi, and R. A. Cole, "Evaluation and Integration of Neural-Network Training Techniques for Continuous Digit Recognition". *Proceedings of International Conference on Spoken Language Processing (ICSLP '98)*, Sydney, Australia, 30 Nov.-4 Dec., 1998, Vol. 3, pp. 731-734.

[3] 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.

[4] 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.

[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] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech". *Journal of the Acoustical Society of America (JASA)*, April 1990, Vol. 87, no. 4, pp. 1738-1752.

[7] 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.

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

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

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

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

[16] B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter and M. Omologo, "Speaker Independent Continuous Speech Recognition Using an Acoustic-Phonetic Italian Corpus". *In Proceedings of International Conference on Spoken Language Processing (ICSLP '94)*, Yokohama Japan, 1994, Vol. 3, pp. 1391-1394.

[17] on the World Wide Web. 1998. European Language Resources Association: http://www.icp.grenet.fr/ELRA/cata/spee_det.html#apasci