CSLU SPEECH TOOLKIT IN ITALIANO: STATO DELL'ARTE

Piero Cosi* - John-Paul Hosom**

*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/

1. SOMMARIO

In questo lavoro sono descritti gli ultimi sviluppi relativi alla realizzazione dei CSLU Speech Toolkit in Italiano e sono illustrati i risultati ottenuti in due specifiche applicazioni di riconoscimento automatico per l'italiano.

In particolare, si fa riferimento allo sviluppo di un sistema di riconoscimento di cifre numeriche connesse "*speaker independent*" (*SI*) su canale telefonico e di un sistema di riconoscimento *SI* "*general purpose*", in cui si devono poter riconoscere tutte le possibili sequenze di parole, su canale microfonico.

2. INTRODUZIONE

Il riconoscimento di sequenze di cifre connesse è importante per moltissime applicazioni telefoniche quali ad esempio: l'accesso telefonico alle informazioni relative al proprio conto corrente o alla propria carta di credito, oppure le chiamate interurbane assistite dal calcolatore. D'altra parte, vi sono molte applicazioni caratterizzate, invece, dal fatto che il lessico non può essere forzatamente limitato, ma deve per forza di cose essere lasciato completamente "libero", conseguentemente, deve essere possibile riconoscere qualsiasi sequenza di parole.

Nella prima applicazione è richiesto un elevatissimo livello d'accuratezza, mentre nella seconda, perdendo il vantaggio di poter utilizzare una qualsiasi grammatica, si rende il processo di riconoscimento strettamente dipendente dall'accuratezza con cui il messaggio verbale è decodificato a livello acustico-fonetico.

Entrambe queste applicazioni richiedono un continuo sviluppo di sempre più aggiornate innovazioni, sia per quanto riguarda l'elaborazione acustica del segnale verbale, sia per quanto riguarda la progettazione di nuove architetture di riconoscimento.

In precedenti lavori, relativi al riconoscimento automatico di cifre connesse, in banda telefonica per l'inglese e su canale microfonico per l'italiano [1-4] sono state ottenute elevate prestazioni. Sono stati realizzati numerosi esperimenti riguardanti il tipo di "*features*" da utilizzare in ingresso al classificatore, che nel nostro caso era costituito da una rete neurale, il tipo di categorie dipendenti-dal-contesto prodotte in uscita dal classificatore stesso, i modelli di durata e la particolare grammatica [4]. Sono stati paragonati i risultati ottenuti mediante sistemi standard basati su HMM e sistemi basati sulla più recente

tecnologia “ibrida” NN/HMM, entrambi implementati mediante i CSLU Speech Toolkit [5], e si è giunti alla conclusione che gli ultimi sono da preferirsi, in termini di prestazioni.

3. CSLU SPEECH TOOLKIT

La piattaforma utilizzata in questo lavoro, è stata, come per i precedenti lavori, il sistema denominato CSLU Speech Toolkit [5], che è disponibile liberamente in rete, per scopi didattici o di ricerca ed include programmi specifici per l’elaborazione del segnale vocale, per il riconoscimento automatico, per la sintesi automatica da testo scritto, per l’animazione di agenti o “facce” parlanti e per la progettazione di sistemi di dialogo.

Questo sistema utilizza, nella sua architettura di base, un approccio “*frame-based*” come illustrato in Figura 1 e 2. Il segnale è diviso in frame e per ogni frame è calcolato un vettore di coefficienti, che ne rappresenta l'inviluppo spettrale, opportunamente pesato per meglio evidenziare alcuni rilevanti aspetti percettivi [6]. [7], e la sua corrispondente energia in una finestra d’analisi di lunghezza prefissata. Questo vettore è calcolato ogni 10 ms. ed è inviato all’ingresso del classificatore neurale. La rete neurale riceve in ingresso non solo il vettore di coefficienti corrispondente al frame in esame, ma anche quello corrispondente ad un numero prefissato di frame adiacenti. Questa “finestra contestuale” di vettori di coefficienti è utilizzata al fine di fornire al sistema informazioni specifiche sulla dinamica del segnale verbale. Ad ogni frame la rete neurale “classifica” l’insieme dei vettori in ingresso in categorie fonetiche, stimando, in pratica, la probabilità che ogni categoria sia rappresentata da quell’insieme di vettori. Il risultato di quest’elaborazione è una matrice CxF di probabilità, dove C è il numero di categorie fonetiche, e F è il numero di frame.

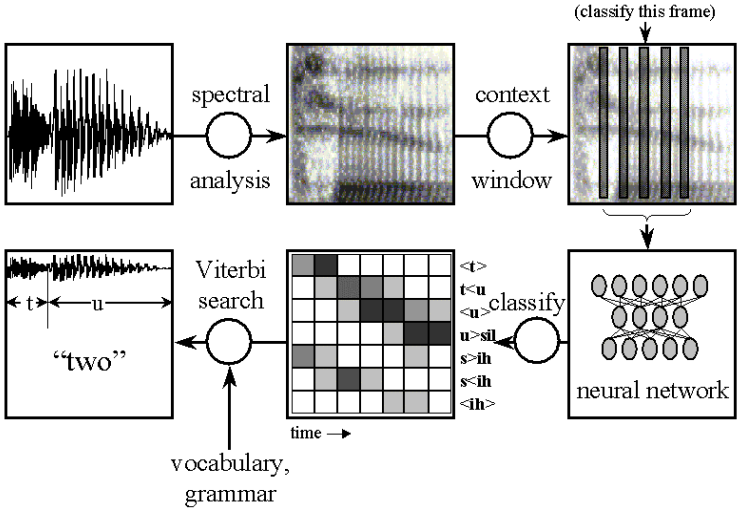

Figura 1. Illustrazione grafica della procedura di riconoscimento del sistema di base del software CSLU Speech Toolkit, relativo alla parola inglese “two”.

La parola, o le parole, che meglio si adattano a questa matrice di probabilità sono determinate mediante un algoritmo di ricerca di Viterbi opportunamente vincolato al particolare vocabolario e alla particolare grammatica utilizzata a seconda delle applicazioni. La ricerca è generalmente considerata come l’attraversamento di una sequenza di stati

(illustrati, in Figura 2, per un semplice vocabolario di 2 parole), dove ogni stato rappresenta una particolare categoria fonetica e sono presenti inoltre le probabilità di transizione da uno stato all'altro.

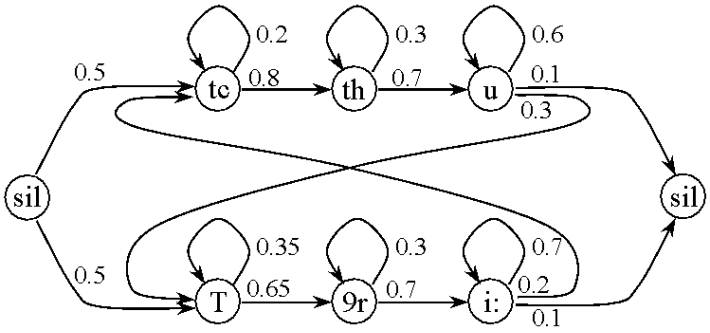

Figura 2. Sequenza di stati HMM per un vocabolario di due parole.

La differenza maggiore di quest'approccio con i sistemi standard basati su HMM risiede nel fatto che le probabilità delle classi fonetiche sono stimate mediante una rete neurale artificiale invece che un insieme di gaussiane, si parla, infatti, di architetture ibride NN/HMM, come illustrato in Figura 3.

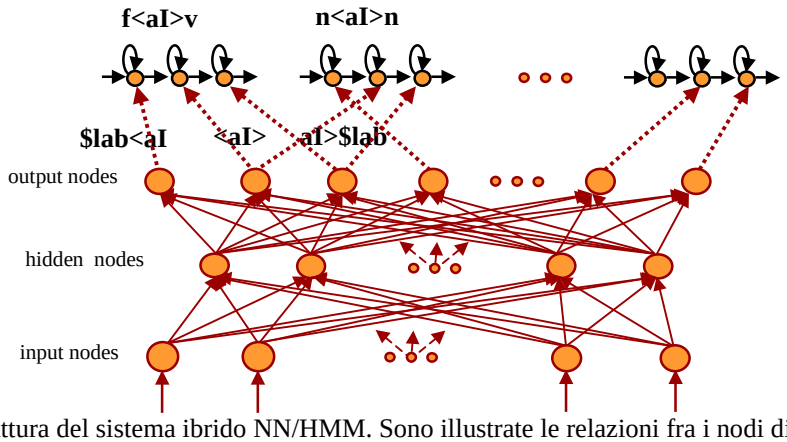

Figura 3. Struttura del sistema ibrido NN/HMM. Sono illustrate le relazioni fra i nodi di uscita della rete neurale artificiale e gli stati dei modelli HMM delle categorie da riconoscere.

Così facendo, non sono necessarie le usuali assunzioni circa la distribuzione statistica e l'indipendenza dei dati in ingresso. Le reti neurali, inoltre, mettono in pratica in modo assai semplice quello che viene denominato "apprendimento discriminativo" ("*discriminative training*" [8]), e la procedura di riconoscimento è molto più veloce di quella realizzabile con i sistemi standard basati su HMM. Un'altra differenza risiede nel tipo di unità fonetiche dipendenti dal contesto. Infatti, mentre nei sistemi basati su HMM l'apprendimento standard viene effettuato considerando i fonemi precedenti o successivi al fonema in esame, nel sistema implementato mediante i CSLU Speech Toolkit ogni fonema

viene diviso in stati che possono essere dipendenti dal contesto destro, sinistro o anche essere totalmente indipendente dal contesto.

4. CIFRE NUMERICHE CONNESSE

Per l'addestramento del sistema di riconoscimento di cifre numeriche per l'italiano su canale telefonico, sono stati utilizzati due corpus denominati FIELD e PHONE [9], entrambe gentilmente resi disponibili dall' IRST in base ad un accordo bilaterale di ricerca. Questi due corpus sono stati entrambe trascritti, a livello di parola e a livello fonetico, mediante una procedura automatica sviluppata dall' IRST e successivamente sono stati manualmente controllati e corretti. La nostra intenzione è stata quella di "addestrare" un sistema sulla base di questi due corpus per poi effettuare la valutazione finale sul corpus telefonico PANDA, reso disponibile dallo CSELT [10]. PANDA costituisce un test estremamente severo essendo in esso contenute sequenze di cifre numeriche, corrispondenti ad altrettanti numeri di carte di credito pronunciate su canale telefonico, di lunghezza media di 16 numeri. Il test di valutazione finale è stato effettuato sul corpus PANDA per dimostrare che il sistema non è "accordato" soltanto ai corpus utilizzati in fase d'addestramento. Il corpus FIELD contiene sequenze di cifre numeriche, corrispondenti ad altrettanti numeri telefonici, raccolte in parte durante un servizio relativo ad un sistema semiautomatico di chiamate telefoniche con operatore, mentre il corpus PHONE contiene sequenze di cifre numeriche casuali raccolte mediante alcune chiamate telefoniche di utenti preventivamente addestrati. Il corpus PHONE è caratterizzato da un notevole numero d'esitazioni, respiri, soffi, colpi di tosse ed altri fenomeni spontanei ed è stato quindi suddiviso in tre sotto-categorie ("*high*, *medium*, e *low*) in base al livello di tali fenomeni spontanei presente nelle varie registrazioni [9]. Poiché la sezione "*low-quality*" contiene principalmente parole non contenute nel vocabolario in oggetto ("cifre numeriche") e la nostra valutazione è concentrata sulle parole all'interno di questo vocabolario, questa sezione non è stata presa in considerazione.

Per quanto riguarda l'analisi acustica, si è considerato un vettore di analisi composto da 26 elementi (13 coefficienti MFCC [7] più il corrispondente "delta") calcolato ogni 10 ms. . Come indicato in Tabella 1, il sistema utilizza delle unità di riconoscimento corrispondenti all'incirca ai fonemi.

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

Tabella 1. Unità acustiche e loro suddivisione in termini di numero di parti (* r significa "dipendente dal contesto destro").

Ogni fonema è dipendente dal contesto relativo a classi fonetiche raggruppate in base al particolare modo/luogo di articolazione, quali ad esempio “vocali posteriori”, “affricate” ecc, come indicato in Tabella 2.

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

Tabella 2. Raggruppamenti di unita acustiche in gruppi di unità simili.

Sono state poi incluse speciali categorie, indipendenti dal contesto, relative alle pause, al rumore dovuto a respiri o soffi ed alle esitazioni. Questa struttura ha dato luogo a poco meno di 100 categorie d’uscita e quindi ad un numero relativamente grande di campioni per ogni classe. La rete neurale risultante, addestrata mediante l’algoritmo di “*back-propagation*”, è una rete neurale a tre livelli, del tipo *feed-forward* interamente connessa ed è costituita da 130 nodi di ingresso, corrispondenti ad una “finestra-contestuale” di 5 frame centrati sul frame target da riconoscere, 200 nodi intermedi e 116 nodi di uscita corrispondenti alle categorie dipendenti dal contesto da riconoscere (Fig. 3).

L’addestramento è effettuato in tre fasi. In un primo momento, utilizzando le trascrizioni fonetiche manuali (*Hand-labeled* training - HL) utilizzando quindi, per la rete neurale, valori di target binari, successivamente utilizzando le trascrizioni fornite automaticamente dal primo stadio, sempre con valori di target binari (*Forced-Alignment* training - FA), ed infine utilizzando le trascrizioni automatiche generate dal secondo stadio questa volta mediante valori di target probabilistici (*Forward-Backward* training - FB). Come illustrato in Figura 4, ad ogni stadio, la valutazione del sistema è stata effettuata su un set di sviluppo (“*development set*”), disgiunto dal set di addestramento, per ogni corpus. Al completamento della fase di sviluppo, la valutazione del sistema è effettuata su un altro set disgiunto di test (“*test set*”).

Attualmente sono state ultimate tutte tre le fasi d’addestramento anche se è stato effettuato un solo ciclo della procedura *Forward-Backward* mentre probabilmente, per poter ottenere un effettivo miglioramento delle prestazioni, ne sono necessari di più. Come illustrato in Tabella 3, infatti, la miglior architettura ottenuta, sia per il corpus FIELD sia per il corpus PHONE, è risultata quella corrispondente alla iterazione 42 della procedura FA, poiché dopo la procedura FB le percentuali di corretto riconoscimento sono risultate lievemente inferiori. In particolare, nell’esperimento di test su FIELD, si sono ottenuti dei valori di corretto riconoscimento a livello di parola (“*Word-evel* A*ccuracy”* - WA) del 99.75% e a livello di frase (“S*entence-level* A*ccuracy”* - SA) del 97.73%, mentre per quanto riguarda PHONE si sono ottenuti dei valori di 98.68% WA e 95.19% SA.

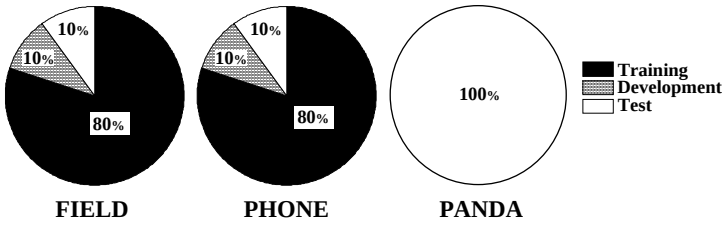

Figure 4. *Training*, *Development* e *Test* set per FIELD (*F*), PHONE (*P*) e PANDA (*PA*). In particolare, 721 sequenze (***s)*** per 6790 cifre numeriche (***c***) in *F* e 1842 ***s*** (7504 ***c***) in *P* sono state usate per il training. 85 ***s*** (791 ***c***) in *F* e 191 ***s*** (791***c***) in *P* sono state usate per il *development*. 88 ***s*** (809 ***c***) in *F*, 208 ***s*** (836 ***c***) in *P* e 1041 ***s*** (16247 ***c***) in *PA* sono state usate per il test.

		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		

Tabella 3. Risultati in termini di percentuali di corretto riconoscimento di parola ("*Word Accuracy*" - WA) e di frase ("Sentence Accuracy" - SA) per la miglior rete ottenuta addestrando il sistema con i dati trascritti manualmente ("*Hand-Labeled"* - HL), ottenuta in seguito all'applicazione della procedura di allineamento-forzato ("*Forced-Alignment* - FA) e di "*Forward-Backward"* (FB). La rete migliore su cui è stato effettuato il test finale del sistema è risultata attualmente la rete migliore ottenuta dopo la procedura di allineamento-forzato in quanto le percentuali di corretto riconoscimento ottenute dopo l'applicazione di un solo ciclo delle procedura FB si sono rivelate lievemente inferiori.

Questi percentuali di riconoscimento sono risultate notevolmente superiori a quelle ottenuti per l'inglese su un analogo corpus di cifre connesse in ambiente telefonico denominato "*CSLU 30Knumbers*" corrispondenti attualmente attorno al 98% e costituiscono per il momento le migliori ottenute su questi dati, come riportato dall' IRST [11] e da CSELT [12]. A livello di parola, infatti, essi corrispondono al 92% e al 71% di riduzione dell'errore di corretto riconoscimento, se paragonati a quelli ottenuti dall' IRST rispettivamente sul corpus FIELD (96.8%) e PHONE (95.5%) [11]. Relativamente ai risultati conseguiti da CSELT, si è riscontrato invece una riduzione dell'errore rispettivamente del 90% e 72% relativamente al corpus FIELD (97.4%) e al corpus PHONE (95.2%) [12].

Anche senza aver terminato completamente la fase di sviluppo, in quanto come già accennato sono forse necessari alcuni cicli in più della procedura FB, è stato effettuato un test finale sul corpus PANDA utilizzando la configurazione migliore ottenuta mediante la

rete risultante dalla procedura FA di *allineamento-forzato*. Questo test ha fornito un'accuratezza a livello di parola del 98.6% e a livello di frase dell' 84.82%, risultato questo comparabile con quello ottenuto da CSELT con il loro miglior sistema [12].

Il test finale sul corpus PANDA, su materiale vocale quindi da considerarsi totalmente disomogeneo rispetto a quello utilizzato in fase di addestramento del sistema, ha fornito una percentuale di corretto riconoscimento a livello di parola del 98.6% e a livello di frase dell' 84.82%. A livello di parola questi valori rappresentano una riduzione dell'errore del 53% rispetto alle migliori percentuali ottenute sullo stesso materiale di test dall' IRST (97.0%) [11], mentre rappresentano invece un aumento dell'errore del 55% se si considera la miglior prestazione ottenuta da CSELT sugli stessi dati (99.1%) [12]. Nel caso degli esperimenti condotti da CSELT vi è però da sottolineare che il materiale utilizzato per l'addestramento del sistema è totalmente incomparabile con quello utilizzato in questo lavoro, sia in termini di "quantità" (8539 sequenze di cifre numeriche corrispondenti ad altrettanti numeri di "carte di credito" facenti parte dello stesso corpus PANDA utilizzate per il training [12]), sia in termini di qualità, in quanto il materiale vocale utilizzato per l'addestramento del sistema appartiene, nel loro caso, allo stesso dominio applicativo di quello utilizzato in fase di test finale.

5. "GENERAL PURPOSE"

Sebbene il riconoscimento di cifre numeriche sia di notevole importanza, in molte altre applicazioni di riconoscimento indipendente dal parlante su vocabolari a dominio specifico (ad esempio "*collect call*", "chiamate con operatore", "accesso a banche dati", ecc.) è necessario disporre di un sistema di riconoscimento "*general-purpose*" in grado di riconoscere tutte le possibili sequenze di fonemi consentite in una determinata lingua. Per sviluppare tale sistema è stato utilizzato il corpus dell' IRST denominato APASCI distribuito dall' ELRA [13]. Questo corpus contiene circa 4000 frasi lette da oltre 150 parlanti, dove le frasi sono state progettate allo scopo di massimizzare il numero di fonemi occorrenti in ogni contesto. ELRA distribuisce questo corpus assieme alle corrispondenti trascrizioni a livello di parola e a livello fonetico, quindi APASCI è stato utilizzato per addestrare un sistema di riconoscimento *general purpose* per l'italiano sulla base di queste trascrizioni. In particolare, 1250 frasi sono state utilizzate per il *training*, 105 per il *development* e 715 per il *test* del sistema, come illustrato in Figura 5.

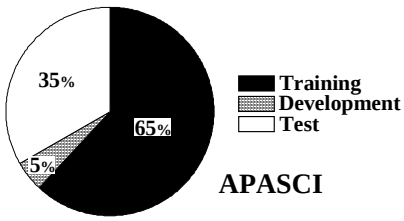

Figura 5. *Training*, *Development* e *Test* set per APASCI In particolare, 1250 frasi per il *training*, 105 per il *development* e 715 per il *test*.

Questo sistema, che utilizza specifiche categorie fonetiche dipendenti dal contesto per tenere in considerazione le variazioni dovute alla coarticolazione, riconosce 38

differenti fonemi o unità acustico-fonetiche (escludendo il silenzio o l'occlusione delle consonanti occlusive) e può distinguere fra le vocali accentate o meno e fra le vocali E ed O aperte o chiuse. L'addestramento è stato sempre progettato, come per le cifre numeriche, in tre fasi e la valutazione del sistema è eseguita in ogni stadio su un insieme di sviluppo ("*development set*")di 105 frasi. Attualmente sono stati completati solo i primi due stadi dell'addestramento ed il livello di accuratezza finale del sistema, come illustrato in Tabella 2, ha raggiunto il valore di 82.90% sul *developmen*t-set di 80.53 sul *test-set*.

	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**

Tabella 4. Percentuali di corretto riconoscimento per APASCI.

Questo livello di accuratezza è assai più elevato di quello ottenuto con lo stesso sistema per un analogo corpus per l'inglese (70%) ed è paragonabile a quello ottenuto dall' IRST sullo stesso corpus APASCI [14], utilizzando un sofisticato sistema HMM (82.44%), e rappresenta una piccola riduzione dell'errore dello 0.8%. Il terzo stadio dell'addestramento relativo alla procedura di *forward-backward* è tuttora in fase di sviluppo e, quando sarà terminato, si procederà alla valutazione finale del sistema sull'insieme di test.

6. CONCLUSIONI

Sono stati realizzati per l'italiano: un sistema di riconoscimento di cifre numeriche su canale telefonico le cui prestazioni rappresentano lo stato dell'arte per l'italiano, ed un sistema di notevoli ed accettabili prestazioni per il riconoscimento fonetico "*general puprose*" su canale microfonico.

Entrambe questi riconoscitori sono stati inclusi nel sistema di dialogo disponibile con i CSLU Speech Toolkit, e sono stati implementati all'interno del sistema due semplici programmi dimostrativi che accettano in ingresso sequenze di cifre numeriche oppure alcuni ordini da menù.

7. SVILUPPI FUTURI

In previsione d'ulteriori sviluppi, è stato progettato, realizzato e testato con successo, inoltre, un nuovo software, che sarà incluso nella prossima versione dei CSLU Speech Toolkit, per consentire la raccolta e la registrazione di materiale vocale direttamente attraverso il canale telefonico, che consentirà all'utente di raccogliere corpus vocali più estesi sui quali sviluppare le proprie ricerche. E' stato realizzato, inoltre, un nuovo software, in grado di consentire all'utente di utilizzare nuove modalità di analisi del segnale vocale, che possono richiedere una grande quantità di tempo di calcolo, quali ad esempio le tecniche basate su modelli uditivi, e di integrarle facilmente con l'attuale procedura di sviluppo di un sistema di riconoscimento, per consentire facili e semplici confronti con le tecniche di analisi standard.

8. RINGRAZIAMENTI

Gli autori ringraziano sinceramente IRST e CSELT per la collaborazione nel rendere disponibili i corpus FIELD, PHONE e PANDA. In particolare, Gianni Lazzari, Daniele Falavigna, Roberto Gretter e Maurizio Omologo dell' IRST e Roberto Billi e Luciano Fissore dello CSELT per il loro supporto e per i loro sempre utili suggerimenti e consigli. Questo lavoro è stato supportato in parte dal Consiglio Nazionale delle Ricerche mediante il programma "*International Short-Term Mobility Program*".

BIBLIOGRAFIA

[1] Cosi P., Hosom J. P., Shalkwyk J., Sutton S., and Cole R.A., *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] Hosom J.P., Cosi P., and Cole R.A., *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] Cosi P. and Hosom J.P., *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] Hosom J.P., Cole R.A., and Cosi P., *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] 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] 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.

[7] 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.

[8] Bourlard H., *Towards Increasing Speech Recognition Error Rates*, Eurospeech'95, vol. 2, pp. 883-894, Madrid, Spain, September 1995.

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

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

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

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

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

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