Associazione Italiana di Acustica
XXVII Convegno Nazionale
Genova, 26-28 maggio 1999

UN SISTEMA IBRIDO (HMM/NN) PER IL RICONOSCIMENTO DI SEQUENZE DI CIFRE CONNESSE IN ITALIANO

P. Cosi [(1)]

[(1)] 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/

SOMMARIO

Viene descritto un sistema per il riconoscimento automatico di sequenze di cifre connesse, indipendente dal parlante, realizzato per l'italiano. Per lo sviluppo e l'implementazione del sistema, è stato utilizzato l'ambiente software denominato ***CSLU-Toolkit,*** mentre l'addestramento e la verifica sono stati resi possibili mediante l'utilizzazione dei corpora denominati ***SPK*** e ***PANDA***. Il sistema si basa su un'architettura "*ibrida*" realizzata mediante catene di Markov nascoste e reti neurali artificiali. Una rete neurale *multistrato* è stata addestrata sulle probabilità di ogni frame di appartenere ad una determinata categoria fonetica. Questa rete è stata poi utilizzata per stimare la probabilità di emissione dei singoli stati di apposite catene di Markov associate alle varie categorie fonetiche e, successivamente, è stato applicato l'algoritmo '*forward-backward*' per stimare nuovi valori target da utilizzare come nuovi *pattern* di addestramento per la rete neurale.

INTRODUZIONE

Lo sviluppo di un sistema di riconoscimento del linguaggio parlato è un'attività molto complessa che generalmente richiede, per la progettazione, la validazione e la vera e propria implementazione del sistema, un lungo periodo che può facilmente durare parecchi mesi o meglio alcuni anni. Sin dai primi anni 90', il "*Center for Spoken Language Understanding"* (***CSLU***)[1] ha lavorato allo sviluppo di nuovi *tool* per facilitare la creazione di sistemi di riconoscimento del linguaggio parlato. Il risultato di questi sforzi si è concretizzato nella realizzazione di un pacchetto software integrato denominato *CSLU-Toolkit* [1] che rappresenta lo stato dell'arte per quanto riguarda la ricerca, lo sviluppo e l'implementazione di sistemi di riconoscimento del linguaggio

[1] CSLU, *Center for Spoken Language Understanding* - OGI, Oregon Graduate Institute of Science and Technology, P.O. Box 91000, Portland Oregon 97291-1000 USA, http://cslu.cse.ogi.edu

parlato[2]. Mediante l'ausilio dei CSLU-Toolkit un numero sempre crescente di utilizzatori, anche non necessariamente esperti nel campo delle tecnologie vocali, può realizzare dei semplici prototipi applicativi di sistemi di riconoscimento, mentre, ai ricercatori esperti è invece fornito un tool efficacissimo per la realizzazione e la validazione delle proprie ricerche, anche le più all'avanguardia nel campo delle tecnologie vocali [2-3].

In questo lavoro è descritto un sistema sviluppato mediante i CSLU-Toolkit per il riconoscimento automatico di sequenze di cifre connesse, indipendente dal parlante, realizzato per l'italiano e sono illustrati i risultati ottenuti sui corpora denominati ***SPK*** [4] e ***PANDA*** [5] progettati e raccolti rispettivamente dall'IRST[3] e dallo CSELT[4].

ARCHITETTURA DEL SISTEMA

Il sistema utilizza l'architettura di base dell'approccio "*frame-based*" dei CLSU-Toolkit illustrato in Figura 1. Il segnale è diviso in *frame* ogni 10 ms e, per ogni frame, è calcolato un vettore combinazione di 13 coefficienti ***PLP*** (*Perceptual Linear Predictive)* [6] e di 13 coefficienti ***MFC*** (*Mel Frequency Cepstral*) [7]. Per cercare di normalizzare e quindi di ridurre l'effetto indotto sul segnale dal diverso canale di trasmissione e dalle diverse caratteristiche dei parlanti, il segnale è preventivamente pre-processato nei due casi, rispettivamente utilizzando la tecnica denominata ***RASTA*** (RelAtive SpecTrAl analysis) [8] e la tecnica denominata ***CMS*** (Cepstral Mean Subtraction) [9].

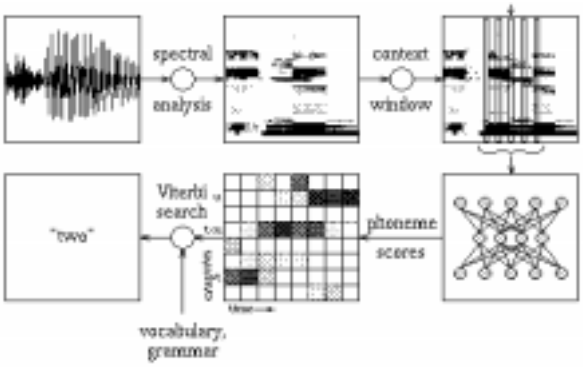

Figura 1. Schema dell'architettura base "*frame-based*" del sistema di riconoscimento implementato dai CSLU-Toolkit.

Il vettore di analisi rappresentante l'inviluppo spettrale del segnale verbale nel frame in esame e nei frame adiacenti rispettivamente a -60, -30, 30, e 60 msec, di dimensione quindi 130, è classificato in 116 categorie fonetiche mediante una rete neurale artificiale a tre livelli, dove nel livello intermedio sono stati considerati 200 nodi. Queste categorie sono dipendenti dal contesto e sono determinate dal lessico di pronuncia delle parole da riconoscere, come indicato in Tabella 1 assieme alla grammatica corrispondente, dal numero di parti con cui dividere ogni unità fonetica a seconda della sua lunghezza e della possibilità che l'unità stessa possa essere più o meno influenzata da fenomeni coarticolatori, e dal tipo scelto di raggruppamento in particolari *cluster* di fonemi simili, come illustrato in Tabella 2. Fonemi composti da due parti

[2] I CSLU-Toolkit sono liberamente recuperabili in rete all'indizizzo internet http://cslu.cse.ogi.edu/toolkit.
[3] IRST – Istituto per la Ricerca Scientifica e Tecnologica- Pantè di Povo, Trento Italy.
[4] CSELT – Centro Studi e Laboratori per le Telecomunicazioni, Torino, Italy.

hanno la parte sinistra dipendente dal fonema precedente e quella destra dipendente dal fonema successivo, mentre nei fonemi di tre parti la parte centrale è considerata più stabile ed indipendente dai fonemi adiacenti. Le uscite della rete neurale sono utilizzate come stime della probabilità, per ognuna di queste categorie fonetiche, che il frame in esame appartenga ad una determinata categoria, e la matrice delle probabilità assieme al lessico di pronuncia sono poi utilizzati all'interno di un algoritmo di ricerca di Viterbi per determinare la sequenza delle parole più probabili.

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}

word	pronunciation
sei	{s E I}
sette	{s E tt e}
otto	{O tt o}
nove	{n O v e}
separ	{.pau [.garbage] .pau}

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

Tabella 1. Lessico e grammatica per le sequenze di cifre dell'Italiano.

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

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 2. Unità fonetiche, numero di parti per ogni unità e raggruppamenti di unità simili.

L'addestramento e la verifica del sistema sono stati effettuati utilizzando una versione filtrata (300-3700 Hz) di SPK [4]. Un ulteriore test finale del sistema è stato effettuato su un sottoinsieme di PANDA [5]. In particolare, per l'addestramento del sistema, sono state utilizzate 20 ripetizioni delle dieci cifre dell'italiano, pronunciate isolatamente, e 20 differenti sequenze di otto cifre connesse, selezionate in modo casuale, appartenenti a 40 parlatori (19 femmine e 21 maschi) quasi tutti originari del Nord Est dell'Italia. Per l'ulteriore test finale del sistema sono state utilizzate delle sequenze di 15 o 16 cifre connesse registrate su canale telefonico [10]. Il segnale è stato acquisito a 48 kHz e 16-bit di accuratezza e sottocampionato a 16 kHz. La segmentazione e la corrispondente trascrizione fonetica sono disponibili soltanto per 10 parlatori, che sono stati quindi utilizzati per l'addestramento del sistema di base (***baseline***). L'addestramento della rete è stato fatto per 30 iterazioni e la rete neurale corrispondente all'iterazione che ha fornito i risultati migliori, indicati in Tabella 3, ottenuti testando il sistema sul segnale corrispondente ai rimanenti 30 parlatori, è stata scelta come la rete neurale di base (***B***). Successivamente, mediante questa rete neurale, il materiale vocale corrispondente a tutti i 40 parlatori è stato forzatamente riallineato (***forced alignment***) per ottenere il nuovo materiale vocale su cui riaddestrare il sistema che è stato poi testato utilizzando questa volta il materiale vocale telefonico vero e proprio contenuto in PANDA. Infine, la nuova rete corrispondente all'iterazione che ha fornito i risultati migliori dopo il riallineamento forzato (***FA***), sempre indicati in Tabella 3, è stata utilizzata per stimare le probabilità di emissione dei singoli stati di apposite catene di Markov associate alle varie categorie fonetiche, come illustrato nella Figura 2. Successivamente, è stato applicato l'algoritmo '*forward-backward*' [10] per stimare

nuovi valori *target* da utilizzare come nuovi *pattern* di addestramento per la rete neurale, che è stata a sua volta testata (***FB***) sullo stesso database telefonico PANDA precedentemente introdotto, ottenendo i risultati sempre indicati nella Tabella 3.

	Itr #	Snts #	Wrds #	Sub %	Ins %	Del %	WrdAcc %	SntAcc %
B	29	4950	8800	0.26	0.10	0.10	99.53	99.29
FA	22	990	15483	2.98	4.31	0.50	92.21	55.15
FB	21	990	15483	2.42	4.56	0.47	92.55	53.74

Tabella 3. Percentuali di corretto riconoscimento a livello di parola (WrdAcc) e di frase (SntAcc) nel caso del sistema di *base* (B), del sistema ottenuto dopo *force alignment* (FA) e del sistema finale ibrido (FB). Sono indicati anche l'iterazione che ha fornito i migliori risultati (Itr), il numero di frasi (Snts) e di parole (Wrds) per il test e la percentuale degli errori di sostituzione (Sub), Inserzione (Ins) e cancellazione (Del).

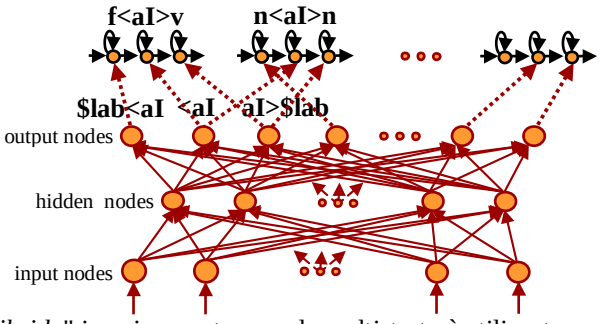

Figura 2. Architettura "*ibrida*" in cui una rete neurale multistrato è utilizzata per la stima delle probabilità di emissione dei singoli stati di apposite catene di Markov associate alle varie categorie fonetiche.

CONCLUSIONI

Le percentuali di corretto riconoscimento nel caso in cui l'addestramento ed il test siano effettuati sullo stesso tipo di materiale vocale (SPK), anche se ovviamente su sottoinsiemi disgiunti, sono ottime sia a livello di parola che di frase (>99%). Il test su materiale telefonico vero e proprio è incoraggiante, anche se a livello di frase il risultato risente ovviamente del *mismatch* fra il tipo di materiale di training e quello di test e potrà essere migliorato utilizzando materiale vocale telefonico anche in fase di training.

BIBLIOGRAFIA

[1] Cole R., Sutton S., Yan Y., Vermeulen P., Fanty M., *Accessible Technology for Interactive Systems: A new approach to spoken language research*, Proc. ICASSP-98, II 1037-1040.

[2] Sutton S., Novick D., Cole R., Fanty M., *Building 10,000 spoken-dialogue systems*, Proc. ICSLP-96, II 709-712.

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

[4] ELRA web page: http://www.icp.grenet.fr/ELRA/cata/spee_det.html#spk

[5] Chesta C., Laface P., Ravera F., *Connected Digit Recognition Using Short and Long Duration Models*. Proc. ICASSP-99, (to be published).

[6] Hermansky H., *Perceptual Linear Predictive (PLP) Analysis of Speech*, JASA, 87- 4, 1738-1752.

[7] Davis S.B., Mermelstein P., *Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences*, IEEE Trans. ASSP, 28- 4, 357-366.

[8] Hermansky H., Morgan N., *RASTA Processing of Speech*, IEEE Trans. SAP, 2-4, 578-589.

[9] Furui S., *Cepstral Analysis Techniques for Automatic Speaker Verification*, IEEE Trans. ASSP, 29-2, 254-272.

[10] Yan Y., Fanty M., Cole R.A., *Speech Recognition Using Neural Networks with Forward-Backward Probability Generated Targets*, Proc. ICASSP-97, 3241-3244.