ESPERIMENTI DI RICONOSCIMENTO FONETICO MEDIANTE RETI NEURALI ARTIFICIALI

P. Cosi

Centro di Studio per le Ricerche di Fonetica - Consiglio Nazionale delle Ricerche
Via Anghinoni, 10 - 35121 Padova (ITALY), Email: cosi@csrf00.csrf.pd.cnr.it

In questo lavoro vengono descritti i principali esperimenti di riconoscimento fonetico mediante Reti Neurali Artificiali effettuati in questi ultimi anni presso il Centro di Studio per le Ricerche di Fonetica del Consiglio Nazionale delle Ricerche di Padova. Dopo una breve panoramica sugli esperimenti di riconoscimento automatico delle vocali, sia inglesi che italiane, e di altre classi fonetiche, verrà descritto più dettagliatamente l'attuale interesse verso le applicazioni di riconoscimento bimodale in cui all'informazione acustica è stata affiancata l'informazione visiva, rappresentata in particolare dalle caratteristiche dinamiche dei movimenti labio-mandibolari, allo scopo di aumentare le prestazioni di riconoscimento soprattutto in ambiente rumoroso.

PHONETIC RECOGNITION EXPERIMENTS BY ARTIFICIAL NEURAL NETWORKS

Some of the main phonetic recognition experiments with Artificial Neural Networks developed during the last years at the *Centro di Studio per le Ricerche di Fonetica* of *Consiglio Nazionale delle Ricerche* will be described. A a very rapid overview of some classification experiments regarding Italian and English vowels and other simple phonetic classes too, will be given. Finally, the present bimodal approach, in which the optic information has been added to the acoustic one, in particular that given by the movements of the lips and the jaw, in order to improve recognition performance principally in a noisy environment, will be illustrated in detail.

1. Introduzione

Le Reti Neurali Artificiali (***RNA***), con particolare riferimento alle Reti Neurali Artificiali Multi-Livello (***RNA-ML***), (***Multi-Level Neural Network***, ***MLNN***), oltre ad essere dal punto di vista teorico delle attraenti strutture elaborative parallele [1], sono un paradigma computazionale molto interessante per l'apprendimento ed il riconoscimento[1] di particolari unità fonetiche o di alcune caratteristiche del segnale verbale [2], [3]. Numerosi algoritmi fra cui quello denominato ***Error Back-Propagation*** (***EBP***) [4], sono stati proposti nel corso degli ultimi anni per consentire a queste strutture di "apprendere" e di "generalizzare" le caratteristiche discriminanti gli stimoli in ingresso, esclusivamente dagli esempi che vengono loro somministrati in fase di apprendimento. Se paragonate ai ben più complessi sistemi in cui ogni conoscenza deve essere esplicitata dallo sperimentatore, è proprio questa caratteristica che rende le RNA molto efficaci in compiti specifici di classificazione di insiemi di dati, quando non vi sia un insieme di regole chiare ed univoche per poterli classificare. In compiti di classificazione fonetica, l'algoritmo EBP consente ad un RNA di apprendere in modo automatico le caratteristiche del segnale vocale in ingresso essenziali per associare o differenziare coppie di *pattern* e possibilmente per creare differenti regioni di decisione per differenti realizzazioni acustiche (ad es. allofoni) della stessa unità fonologica. Contrariamente all'apprendimento dei modelli basati sulla teoria delle catene 'nascoste' di Markov (***Hidden Markov Model, HMM***) [5] che, nella sua applicazione abituale, costruisce un modello per ogni classe senza tenere in considerazione le conoscenze relative alle altre classi, mediante questo paradigma neurale si apprende dalla presentazione successiva di tutti gli esempi in ingresso quello che rende due classi differenti e quello che rende due elementi della stessa classe simili fra loro. Fra le unità del livello di uscita si instaura una competizione per modificare l'influenza delle unità "nascoste" del livello inferiore in modo che queste contribuiscano alla riduzione dell'errore per ogni singola unità di uscita.

Nei primi esperimenti di riconoscimento vocalico si è fatto uso di RNA '***statiche***' in cui l'informazione viene fornita in ingresso al sistema in un unico vettore in cui vengono raccolti tutti i parametri o gli indici di interesse su cui il sistema stesso deve operare per apprendere, mediante EBP, le caratteristiche degli stimoli in ingresso e realizzarne successivamente la relativa classificazione in fase di riconoscimento. Viceversa, negli esperimenti riguardanti classi fonetiche più complesse ed anche in quelli relativi al riconoscimento bimodale si è preferito utilizzare delle strutture '***dinamiche***', denominate Reti Neurali Artificiali Ricorrenti (***RNA-R***) [6], in cui l'informazione viene fornita al sistema in modo dinamico, quindi per ogni finestra di analisi, e la supervisione viene effettuata in più istanti mediante un'estensione degli algoritmi di apprendimento sviluppati per il caso statico denominati *Extended Error Back Propagation for Sequences* (***EEBPS***) [7].

In tutti gli esperimenti, l'analisi spettrale classica (*FFT, LPC, Cepstrum...*) è stata sostituita in termini di *front-end* acustico da un modello del sistema uditivo periferico umano [8] dimostratosi assai efficace soprattutto in condizioni rumorose [9], [10], [11]. Vista la ancora imbattibile abilità umana di percepire univocamente i vari fonemi, anche se

[1] Il termine ***riconoscimento*** è da intendersi, in questa trattazione, sinonimo del termine ***classificazione***.

enormemente modificati dal contesto in cui appaiono, e presupponendo che una prima normalizzazione degli stimoli acustici avvenga a livello del sistema uditivo periferico, un complesso modello di analisi basato sulle più recenti conoscenze neurofisiologiche dell'apparato uditivo periferico umano è sembrato più idoneo ed efficace allo scopo propostoci. Non essendo l'elaborazione acustica oggetto di questa breve rassegna di esperimenti questo modello di analisi non verrà qui descritto e, per una trattazione dettagliata, si rimanda ad alcuni lavori apparsi in letteratura [8], [12].

2. Riconoscimento statico di vocali

Sia per l'inglese americano (**IA**) che per l'italiano (**I**) sono stati progettati alcuni esperimenti di classificazione vocalica. Per quanto riguarda IA [13] sono state considerate le 10 vocali /**i**,**I**, **E**, **@**, **^**, **}**, **a**, **O**, **U**, **u**/ estratte automaticamente mediante un algoritmo di segmentazione [14] dall'insieme delle 10 parole *BEEP, PIT, BED, BAT, BUT, FUR, FAR, SAW, PUT, BOOT* pronunciate 5 volte da 20 parlanti (10 maschi e 10 femmine), mentre, per quanto riguarda I [15], sono state considerate le 7 vocali /**i**, **e**, **E**, **a**, **O**, **o**, **u**/ estratte dall'insieme delle 7 parole *PIPA, PEPE, PEPPA, PAPA, POPE, POPPA, PUPA* pronunciate 5 volte da 10 parlanti maschi. Il front-end acustico del sistema fornisce ogni 5 ms. un vettore di 40 coefficienti [12] contenente "neuralmente" l'informazione spettrale del segnale in ingresso. Per omogeneizzare le informazioni dinamiche del segnale e per semplificare la progettazione della rete è stata applicata un'interpolazione per ridurre ad una costante (IA: c=10, I: c=5) il numero dei *frame* utilizzabili per ogni vocale in ingresso escludendone una percentuale costante della durata totale di ogni stimolo, all'inizio ed alla fine dello stimolo stesso. La rete neurale utilizzata in questi esperimenti, descritta in Figura 1 per il caso I, è una rete in cascata a tre livelli dove ogni livello è completamente connesso con il livello superiore. Vi sono 10 nodi (IA) oppure 7 nodi (I) nel livello di uscita per le vocali in esame, 20 nodi nel livello intermedio (questa è risultata alla luce delle prove effettuate l'architettura più conveniente) e 400 nodi (IA: 10 *frame* x 40 coeff. neurali[2]) oppure 200 nodi (I: 5 *frame* x 40 coeff. neurali) nel livello di ingresso. Come paradigma di apprendimento si è utilizzato l'algoritmo EBP e, per quanto riguarda la procedura di convergenza, è stata utilizzata la tecnica del gradiente nella misura dell'errore quadratico [4]. E' stata adottata la versione "*on-line*" di questo algoritmo per incrementare la velocità di apprendimento, in opposizione alla versione "*batch*" (nella versione on-line l'aggiornamento dei pesi delle connessioni della rete è realizzato dopo la presentazione di ogni pattern in ingresso mentre in quella batch l'aggiornamento dei pesi delle connessioni della rete è realizzato successivamente alla presentazione di tutti i pattern utilizzati per l'apprendimento). La regola di aggiornamento dei pesi delle connessioni include anche un termine di accelerazione utilizzato per smussare ed incrementare ulteriormente la velocità dell'apprendimento stesso. In questi esperimenti di classificazione la percentuale di riconoscimento è stata, nel caso IA, del 95.7% (98.5% e 99.4 se la risposta corretta viene ricercata rispettivamente fra le prime due o le prime tre alternative) e, nel caso I, del 99%, dove in un solo caso la vocale /**o**/ è stata classificata come /**O**/. Come in tutti gli esperimenti che verranno descritti successivamente anche nel caso appena descritto gli stimoli utilizzati in fase di test del sistema, appartengono ovviamente a parlanti non utilizzati in fase di apprendimento. In particolare, per quanto riguarda IA, 13 parlanti (7 maschi e 6 femmine) sono stati considerati per l'apprendimento e 7 (3 maschi e 4 femmine) per il test, mentre nel caso I, per l'apprendimento sono stati considerati 7 parlanti e per il test i rimanenti 3. Per stimoli appartenenti ai parlanti utilizzati per l'apprendimento, e quindi considerando il caso di riconoscimento dipendente dal parlatore, la percentuale ha sempre raggiunto valori elevatissimi vicini al 100% sia nel caso IA che nel caso I. Questi risultati dimostrano un notevole miglioramento se confrontati con quelli di un analogo esperimento effettuato solo per l'inglese americano (85%) in cui al modello del sistema uditivo era stato sostituito un banco di filtri la cui frequenza centrale era logaritmicamente distribuita [13].

Per quanto riguarda l'italiano è stato effettuato anche un altro esperimento di classificazione vocalica [15] su un insieme di vocali pronunciate isolatamente una sola volta da 20 parlanti [16] (14 parlanti sono stati utilizzati per l'apprendimento e 6 per il test) e, anche in questo caso, si è ottenuta un elevata percentuale di classificazioni corrette 98%.

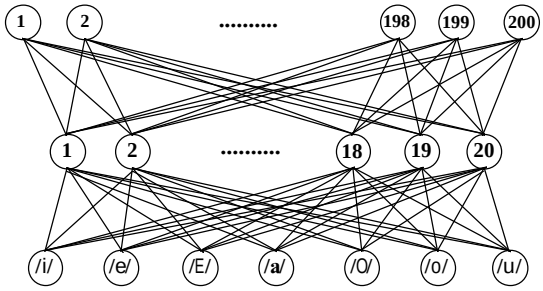
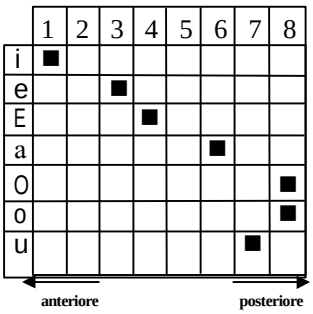


Figura 1. RNA Multi-Livello per il riconoscimento *statico* delle vocali italiane.

Tutti gli esperimenti si qui descritti sono stati poi ripetuti cercando di riconoscere, non le vocali, ma una particolare codifica delle vocali stesse, basata sulla loro classificazione articolatoria relativa al luogo orizzontale e verticale di articolazione, in altre parole al loro grado di *apertura* e *chiusura*. Questo è stato fatto pensando di esplorare successivamente la possibilità di riconoscere sia in modo statico stimoli vocalici non appartenenti all'insieme delle vocali utilizzato per l'apprendimento, per l'inglese americano, oppure, in modo dinamico, alcune sequenze vocaliche come i dittonghi o gli iati per l'italiano. Ad esempio per l'italiano sono state realizzate due RNA denominate LOA e LVA, del tutto analoghe a quella illustrata in Figura 1, alle quali però è stata fatta apprendere automaticamente e sempre in modo statico la codifica descritta in Tabella 1 per le vocali italiane. In altre parole i nodi di uscita non identificano la vocale ma la sua codifica. I risultati ottenuti combinando assieme le risposte delle due reti sono stati analoghi a quelli ottenuti precedentemente con una sola rete [12], [16].

[2] I parametri in uscita dal modulo GSD del modello di analisi uditiva [8] sono 40.

LOA (Luogo Orizzontale di Articolazione)



LVA (Luogo Verticale di Articolazione)

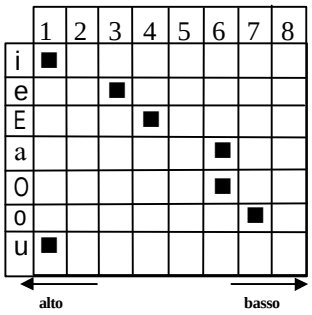


Tabella 1. Codifica delle vocali italiane nel dominio dei luoghi orizzontale e verticale di articolazione (si è voluto codificare le vocali italiane utilizzando gli 8 livelli necessari per la codifica delle vocali inglesi americane in funzione di un possibile confronto fra i due sistemi).

3. Riconoscimento dinamico di sequenze vocaliche

Come accennato nel paragrafo precedente, le due reti denominate LOA e LVA sono state utilizzate con risultati incoraggianti per classificare *dinamicamente* alcune sequenze vocaliche non utilizzate in fase di apprendimento delle reti stesse [12], [17]. Innanzitutto, il termine dinamico si riferisce al fatto che le vocali, o in generale gli stimoli da classificare, non vengono più trattati come un'unica identità rappresentata da un unico vettore, contenente tutte le informazioni sulle loro caratteristiche spettrali, ma vengono trattati in modo dinamico come una successione di *frame*. Scompare quindi l'esigenza di una pre-normalizzazione sulla durata degli stimoli che era necessaria a causa delle dimensioni costanti della RNA in termini di numero di nodi utilizzati per il livello di ingresso. In questo nuovo approccio il vettore di ingresso scorre dinamicamente sullo stimolo e la classificazione viene effettuata con una frequenza costante definita a priori. La finestra di analisi considerata è stata dunque impostata a 10 o 5 ms. rispettivamente nel caso IA e I, ed il passo di analisi è stato impostato a 5 ms. In altre parole le RNA del precedente paragrafo, apprese in modo statico, sono state applicate in modo dinamico a più stimoli vocalici e, anche se non sono state analizzate grosse basi di dati e non si possono trarre considerazioni definitive, i risultati degli esperimenti di classificazione ottenuti sono assai interessanti soprattutto in funzione del parallelo fonetico e articolatorio che può essere effettuato in base ad un'analisi dettagliata delle risposte fornite dalle RNA in esame. In Figura 2 è illustrato un esempio di classificazione di una sequenza vocalica italiana.

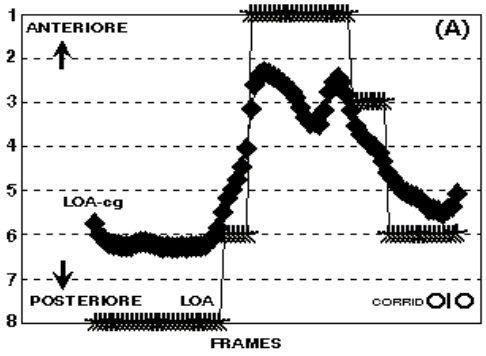


Figura 2. Andamento delle uscite della RNA preposta alla classificazione del luogo orizzontale di articolazione relativamente alla sequenza vocalica / **0j** / estratta dallo stimolo /**koridOj**/. E' illustrato in figura anche l'andamento del corrispondente centro di gravità calcolato con i valori di tutti i nodi di uscita.

Si può facilmente notare come le configurazioni articolatorie dello stimolo in esame, sconosciuto al sistema in fase di apprendimento, siano correttamente classificate sulla base delle conoscenze fono-articolatorie note a priori per le realizzazioni di tali stimolo.

4. Esperimenti di riconoscimento fonetico mediante reti neurali artificiali ricorrenti

Molti tipi di RNA sono stati studiati per risolvere il problema della classificazione automatica di un insieme di dati a partire dai dati stessi. Sia RNA '***statiche'***, come quelle illustrate nel primo paragrafo, in cui lo stimolo è interamente ed in modo univoco presentato in ingresso alla rete, come anche RNA '***dinamiche'***, in cui l'informazione alla rete viene fornita in modo dinamico, come ad esempio quelle illustrate nel paragrafo precedente, hanno fornito risultati interessanti che giustificano ampiamente una loro effettiva utilizzazione in compiti di riconoscimento fonetico, in alternativa alle più classiche tecniche statistiche di *pattern recognition* basate sulla teoria delle HMM [5]. Fra le RNA statiche le RNA-ML [1] addestrate mediante BP sono probabilmente le più diffuse, mentre, per quanto riguarda le RNA dinamiche non c'è ancora un'interpretazione univoca su quale sia il metodo più efficace per introdurre un comportamento dinamico all'interno dei vari sistemi:

(a) la trasformazione di reti ricorrenti in reti di tipo '***feedforward'*** [18];
(b) l'introduzione di connessioni di '***feedback'*** [19];
(b) l'introduzione di buffer contestuali sullo stadio di ingresso [20];
(c) l'introduzione di buffer contestuali sugli stadi di ingresso e su quelli nascosti [21];

sono tutti esempi di come si sia cercato nel corso degli anni di includere un comportamento dinamico che fosse in grado di consentire alle RNA di '*ricordare*' ciò che viene presentato in ingresso, in altre parole di mantenere una memoria, a breve o lungo termine, del passato. Soprattutto, l'ultimo approccio (c), ha fornito recentemente ottimi risultati in termini di prestazioni in sistemi automatici di riconoscimento fonetico interamente basati su RNA [2], [3]. Pur codificando l'informazione in modo dinamico, tutti i metodi sopra elencati risentono di una intrinseca limitatezza nella loro rappresentazione delle informazioni sul passato in quanto ognuno di essi utilizza algoritmi di apprendimento che necessitano sempre di un instante temporale preciso in cui

attivare la supervisione e quindi l'informazione alla rete viene sempre fornita, mediante una modalità statica, cioè dopo aver ultimato la ricezione dello stimolo da classificare. Negli esperimenti che verranno descritti successivamente sono state considerate, invece, delle RNA denominate Reti Neurali Artificiali Dinamiche Multi-Livello (***RNA-DML***) (***Dynamic Multi-Layered Neural Network***, ***DMLNN***) in cui la supervisione viene effettuata in modalità dinamica, mediante un algoritmo denominato EEBPS [7], senza considerare quindi un vettore di ingresso unico in modalità statica. In altre parole, invece di aspettare un istante prefissato, l'algoritmo di apprendimento viene attivato in continuità e la supervisione sui nodi di uscita della rete viene effettuata durante l'evoluzione delle attivazioni. L'ambiente dell'apprendimento viene definito dalla sequenza dei *frame* che rappresentano la naturale evoluzione temporale del segnale in ingresso. Il modello dinamico considerato è discreto, non continuo, e le sue transizioni avvengono quando un nuovo *frame*, e quindi un nuovo vettore di parametri caratterizzanti gli stimoli da classificare, viene applicato in ingresso. Le reti utilizzate sono un caso molto semplice delle RNA-DML, ed in particolare si riferiscono a quelle strutture in cui gli unici neuroni dinamici, quelli cioè che hanno connessioni con feedback su se stessi, sono presenti internamente solo al livello nascosto della rete, quello cioè che riceve le connessioni dal livello di ingresso.

5. Riconoscimento dell'I-set e dell'E-set.

Per valutare l'efficacia delle RNAR quali classificatori automatici sono stati elaborati due esperimenti di riconoscimento [22] di due classi fonetiche italiane particolarmente *complessi* da classificare. La prima classe, denominata ***I-set,*** è costituita dalle consonanti dell'alfabeto italiano articolate con la vocale I (*grafema*), la vocale I, e altri due stimoli sempre articolati con I particolarmente frequenti in Italiano:

/ bi/, / ʧi/, / di/, / ʤi/, / pi/, / ti/, / vi/
+ / i/ + /'i/, / ʃi/;

la seconda classe, denominata ***E-set,*** è costituita dalle consonanti dell'alfabeto italiano articolate con la vocale E (*grafema*),:

/ ɛf:e/, / ɛl:e/, / ɛm :e/, / ɛn:e /, / ɛr:e /, / ɛs:e /.

In questi esperimenti la dimensione del vettore di analisi uditiva fornito dal front-end acustico è di 80 e non più di 40 come negli esperimenti illustrati nei primi paragrafi in quanto non solo l'uscita dal **GSD** (**General Synchrony Detector**), ma anche quella dall'**ED** (**Envelope Detector**) [8] viene presa in considerazione per la classificazione. Vista la natura fortemente *transizionale* degli stimoli appartenenti a queste due classi fonetiche si è ritenuto di dover includere anche i parametri forniti dall'ED che, pur non identificando le caratteristiche spettrali alla pari di quelli forniti dal GSD, forniscono utilissime informazioni sulla natura dinamica degli stimoli in esame. Sono stati utilizzati 7 parlanti maschi adulti di età compresa fra i 19 e i 20 anni e tutti hanno ripetuto 5 volte il materiale vocale per un totale di 350 stimoli per l'I-set e 210 stimoli per l'E-set. Per aumentare la rilevanza statistica dei risultati ogni esperimento di classificazione è stato ripetuto 7 volte utilizzando circolarmente 6 parlanti per la fase di addestramento ed 1 parlante per la fase di test del sistema. Tutti gli stimoli sono stati semi-automaticamente segmentati con SLAM, un nuovo sistema di segmentazione e labelling sviluppato al CSRF [23] che si è dimostrato molto utile ed efficace soprattutto in termini di velocità e precisione anche in condizioni di segnale rumoroso [24]. Si può notare, osservando la struttura delle RNA-R illustrate in Figura 3, come, per entrambe le reti, siano stati considerati 80 nodi di ingresso, corrispondenti alla dimensione del vettore acustico-uditivo fornito dal front-end, mentre, per quanto riguarda il livello nascosto, per l'I-set siano stati considerati 20 nodi dinamici, mentre per l'E-set soltanto 8. Relativamente al livello di uscita sono stati considerati ovviamente 10 nodi statici per l'I-set e 6 per l'E-set. La supervisione è stata effettuata in un solo *frame*, alla fine ed alla metà della consonante target rispettivamente per l'I-set e per l'E-set, per velocizzare i tempi di apprendimento e verificare l'efficacia della classificazione nelle condizioni più critiche possibili. In Tabella 2 sono riassunti i risultati che, considerando la difficoltà del compito in esame, dimostrano la possibilità di utilizzare tali strutture come classificatori fonetici.

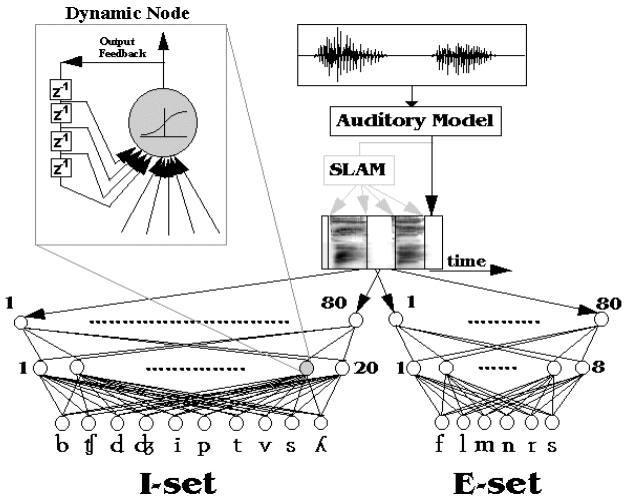


Figura 3. RNA per il riconoscimento dell'Iset e dell' E-set italiani. Sulla destra è illustrata la struttura di uno dei nodi dinamici del livello intermedio.

soggetto	I-set % errore	E-set % errore
MM	22	7
GF	32	30
PT	36	13
SR	48	10
BC	48	10
EP	36	10
MR	27	3
media	**35**	**12**

Tabella 2. Prestazioni degli esperimenti di riconoscimento dell'I-set e dell'E-set italiani (in % di errore).

6. Riconoscimento bimodale delle occlusive

Gli ultimi esperimenti effettuati al CSRF riguardano il riconoscimento bimodale delle occlusive italiane / p, b, t, d, k, g / articolate con le tre vocali cardinali /i, a, u /

[25], [26], [27], [28]. Anche se la modalità uditiva rappresenta il flusso di informazione più importante nel processo di percezione del segnale verbale è ormai appurato che la modalità visiva (***speech-reading, lip-reading***) consente di aumentare la capacità percettiva [29], [30] soprattutto quando il canale uditivo risulta essere fortemente corrotto da rumore [31]. L'ipotesi di base per questi esperimenti è stata che utilizzando i dati relativi ai movimenti articolatori delle labbra e della mandibola dei parlanti, in altre parole simulando le informazioni estratte mediante la lettura labiale, le prestazioni di un sistema di riconoscimento fonetico potessero essere notevolmente migliorate. Utilizzando ***ELITE*** [32], un sistema in grado di registrare in tempo reale i movimenti dinamici di alcuni *marker* riflettenti la luce infrarossa trasmessa e ricevuta da due speciali telecamere, si sono potute simulare tali condizioni di percezione bimodale. Questo esperimento è stato effettuato sia in modalità dipendente che indipendente dal parlatore e sono stati utilizzati 4 soggetti (2 maschi e 2 femmine) nel primo caso e 10 soggetti maschi nel secondo. Ogni soggetto ha pronunciato gli stimoli 5 volte e l'esperimento è stato ripetuto circolarmente 10 volte utilizzando 9 soggetti per l'apprendimento e 1 per il test. In figura 4 è illustrata la posizione dei marker sul volto dei soggetti mentre in Figura 5 è illustrata la struttura dell'intero sistema e della RNA utilizzata che, come si può osservare, è la stessa RNA-DML degli esperimenti descritti nel paragrafo precedente, soltanto che in questo caso vi sono due sezioni parzialmente separate relative: una all'ingresso acustico corrispondente agli stessi parametri uditivi forniti dal front-end precedentemente utilizzato e l'altra all'ingresso visivo corrispondente ai parametri articolatori forniti da ELITE. Questa volta la supervisione è effettuata in due *frame*, ed in particolare a circa metà ed alla fine del segmento occlusivo target, per caratterizzare le due zone in cui l'informazione visiva e quella uditiva sono *maggiormente rilevanti*. Per quanto riguarda gli esperimenti condotti in modalità dipendente dal parlatore, i risultati ottenuti, riassunti nella Tabella 3, indicano chiaramente come l'introduzione dell'informazione visiva abbia portato ad un chiaro miglioramento delle prestazioni anche quando gli stimoli erano fortemente corrotti da rumore (S/N 0dB), mentre, i risultati riassunti in Tabella 4, dimostrano l'efficacia della struttura proposta anche in modalità indipendente dal parlatore.

7. Conclusioni

I risultati ottenuti negli esperimenti di riconoscimento fonetico mediante RNA condotti al CSRF, brevemente illustrati in questo lavoro, hanno ampiamente dimostrato la possibilità di realizzare sistemi di riconoscimento a livello fonetico alternativi alle usuali tecniche statistiche. Gli sviluppi futuri delle ricerche verteranno pincipalmente sul riconoscimento fonetico bimodale nell'ottica di una sempre maggior integrazione di tutte le informazioni disponibili al fine di incrementare le prestazioni del sistema finale. Pur consci dell'attuale impossibilità di costruire sistemi di riconoscimento basati su apparecchiature altamente specializzate quali ELITE, questi studi tendono esclusivamente a dimostrare la validità scientifica dell'approccio bimodale, anche in funzione dell'inevitabile e sempre più veloce sviluppo tecnologico delle interfacce uomo-macchina che consentirà al calcolatore del futuro di dotarsi di organi sensoriali artificiali non solo uditivi, ma anche visivi.

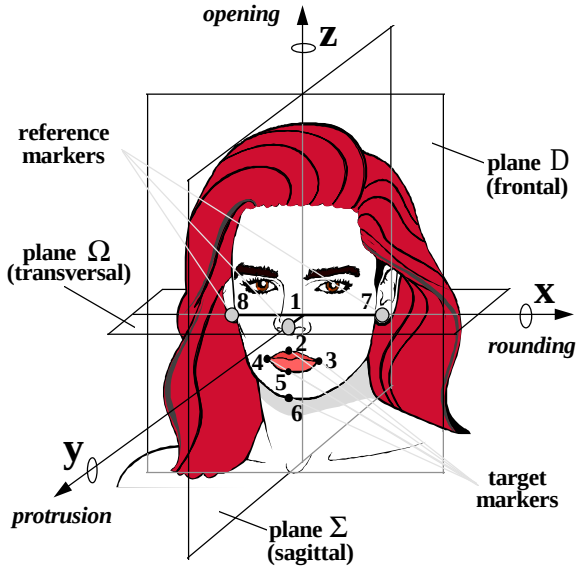


Figura 4. Posizionamento dei marker riflettenti utilizzati con ELITE così come sono stati posizionati sul volto dei soggetti nell'esperimento di riconoscimento delle occlusive dell'Italiano. Sono indicati anche i piani di riferimento per alcuni dei parametri utilizzati per il riconoscimento.

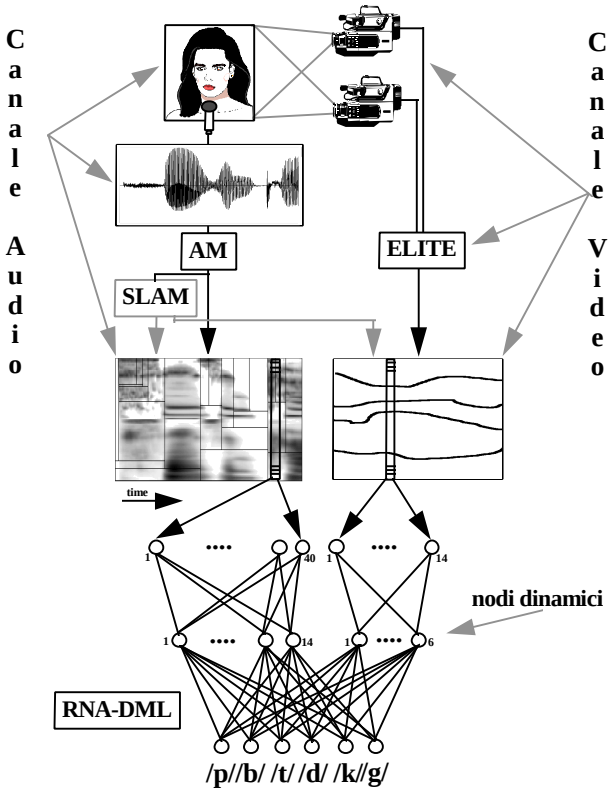


Figura 5. Struttura del sistema di riconoscimento bimodale nell'esperimento di riconoscimento delle occlusive dell'Italiano. Sono indicati anche i piani di riferimento utilizzati nella determinazione di alcuni parametri.

soggetto	AC	AR	AR(PLA)	AC+AR
MA(m)	83	67	100	100
LI(m)	78	61	97	97
PA(f)	78	67	98	96
AN(f)	72	72	100	98
media	**78**	**67**	**99**	**98**

soggetto	AC	AC+AR
MA(m)	83	100
LI(m)	78	94
PA(f)	67	95
AN(f)	67	94
media	**74**	**96**

Tabella 3. Risultati degli esperimenti di riconoscimento delle occlusive in condizioni dipendenti dal parlante. La tabella inferiore si riferisce all'esperimento condotto in condizioni rumorose (S/N 0 dB). Le varie colonne indicano rispettivamente i risultati degli esperimenti condotti utilizzando: solo l'informazione acustica (AC), solo l'informazione articolatoria (AR), solo l'informazione articolatoria ma raggruppando i risultati in termini di classi di articolazione (occlusive labiali, dentali, velari) (AC(PLA)), entrambe le informazioni acustica ed articolatoria AC+AR.

soggetto	% corretto	% corretto PLA
S1	84.4	87.8
S2	83.3	83.3
S3	93.3	93.3
S4	64.4	71.1
S5	47.8	56.7
S6	53.3	64.4
S7	75.6	76.7
S8	60.0	76.7
S9	62.2	71.1
S10	86.7	91.1
media	**71.1**	**77.22**

Tabella 4. Risultati degli esperimenti di riconoscimento delle occlusive in condizioni indipendenti dal parlante. In questo esperimento sono state utilizzate entrambe le informazioni acustica e articolatoria. La colonna di destra si riferisce ai risultati ottenuti raggruppando le singole occlusive in termini di classi di articolazione (occlusive labiali, dentali, velari) .

Bibliografia

[1] D.E. Rumelhart, J.L. McClelland, and the PDP Research Group, *Parallel Distributed Processing: Exploration in the Microstructure of Cognition*, MIT Press, 1986.

[2] J.B. Hampshire II and A. Waibel, *Connectionist Architectures for Multi-Speaker Phoneme Recognition*, CMU-CS-89-167, August 31, 1989.

[3] A.J. Robinson and F. Fallside, *A Recurrent Error Propagation Network Speech Recognition System*, Computer, Speech and Language, Vol. 5, 1991, pp. 257-286.

[4] D.E. Rumelhart, G.E. Hinton and R.J. Williams, *Learning Internal Representations by ErrorPropagation*, in Parallel Distributed Processing: Exploration in the Microstructure of Cognition, Vol. 1, Foundations, MIT Press, 1986, pp. 318-362.

[5] L.R. Rabiner, S.E. Levinson and M.M. Sondhi, *On the Application of Vector Quantization and Hidden Markov Models to Speaker Independent isolated Woed Recognition*, Bell System Technical Journal, Vo. 62, 1983, pp. 1075-1105.

[6] P. Frasconi, *Progetto e realizzazione di un simulatore per reti neurali ricorrenti e sviluppo di prototipi per il riconoscimento vocale in tempo reale*, Tesi di Laurea, Università di Firenze, Dipartimento di Sistemi ed Informatica, Corso di Laurea in Ingegneria Elettronica, Anno Accadenico 1989-1990

[7] M. Gori, Y. Bengio and R. De Mori, *BPS: A Learning Algorithm for Capturing the Dynamical Nature of Speech*, Proceedings of the IEEE-IJCNN89, Washington, June 18-22, 1989, Vol. II, pp. 417-432.

[8] S. Seneff, *A joint synchrony/mean-rate model of auditory speech processing*, Journal of Phonetics, January 1988, pp. 55-76.

[9] M.J. Hunt and C. Lefebvre, *Speaker Dependent and Independent Speech Recognition Experiments with an Auditory Model*, Proceedings IEEE-ICASSP-88, New York, April 11-14, 1988, pp. 215-218.

[10] P. Cosi, *Auditory Modelling for Speech Analysis and Recognition*, in Visual Representation of Speech, M. Cooke, S. Beet and M. Crawford eds., John Wiley & Sons Ltd., 1992, pp. 205-212.

[11] C.R. Jankowski Jr., Hoang-Doan H. Vo and R.P. Lippmann, *A Comparison of Signal Processing Front Ends for Automatic Word Recognition*, IEEE Transactions on Speech and Audio Processing, Vol. 3, N. 4, July 1995, pp. 286-293.

[12] P. Cosi, Y. Bengio and R. De Mori, *Phonetically-Based Multi-Layered Neural Networks for Vowel Classification*, Speech Comm., Vol. 9, N. 1, February 1990, pp. 15-29.

[13] Y. Bengio, R. Cardin, P. Cosi and R. De Mori, *Speech Coding with Multi-Layer Networks*, Proceedings of IEEE ICASSP-89, International Conference on Acoustic Speech and Signal Processing, Glasgow, 23-26 May 1989, pp. 164-167.

[14] R. De Mori, P. Laface and Y. Mong, *Parallel algorithms for syllable recognition in continuous speech*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-7, N. 1, 1985, pp. 56-69.

[15] P. Cosi, R. De Mori and K. Vagges, *A Neural Network Architecture for Italian Vowel Recognition*, Proceedings of VERBA-90, International Conference on Speech Technologies, Rome, 22-24 January, 1990, pp. 221-228.

[16] F. Ferrero, *Diagrammi di esistenza delle vocali Italiane*, Alta Frequenza, 37, 1968, pp.54-58..

[17] P. Cosi, *Riconoscimento di sequenze vocaliche tramite reti neurali*, in, Magno Caldognetto and P. Benincà (Eds.), Interfaccia tra Fonologia e Fonetica, 1991, pp.165-172.

[18] D.C. Plout and G.E. Hinton, *Learning Sets of Filters Using Backpropagation*, Computer Speech and Language, Vol. 2 (2), July 1987, pp. 35-61.

[19] F.J. Pineda, *Generalization of Back-Propagation to Recurrent Neural Networks*, Physical Review Letters, Vol. 59, n. 19, November 1987, pp. 2229-2232.

[20] T.J. Sejnowsky and C.R. Rosemberg, *NETTalk: a Parallel Network that Learns to Read Aloud*, Technical Report JHU/EECS-86/01, 1986.

[21] A. Waibel, T. Hanazawa, G.E. Hinton, K. Shikano and K. Lang, *Phoneme Recognition Using Time-Delayed*

Neural Networks, A.T.R. Technical Report TR-I-0006, October 1987.

[22] P. Cosi, G.A. Mian and M. Contolini, *Speaker Independent Phonetic Recognition Using Auditory Modelling and Recurrent Neural Networks*, Proceedings of ICANN-94, International Conference on Artificial Neural Networks, Sorrento, Italy, 26-29 May, 1994, pp. 925-928.

[23] P. Cosi, *SLAM: a PC-Based Multi-Level Segmentation Tool*, in Speech Recognition and Coding. New Advances and Trends, A.J. Rubio Ayuso and J.M. Lopez Soler edts, NATO ASI Series, Computer and Systems Sciences, Springer Verlag, Vol. F 147, 1995, pp. 124-127.

[24] P. Cosi, *On The Use of Auditory Models in Speech Technology*, in V. Roberto Ed., Lecture Notes in Artificial Intelligence: Intelligent Perception Systems, Springer Verlag Publisher, Vol. 745, 1993, pp. 85-103.

[25] P. Cosi, E. Magno Caldognetto, K. Vagges, G.A. Mian e M. Contolini, *Bimodal Recognition Experiments with Recurrent Neural Networks*, Proceedings of IEEE ICASSP-94, International Conference on Acoustic Speech and Signal Processing, Adelaide. Australia, 19-22 April, 1994, paper 20.8.

[26] P. Cosi, M. Dugatto, F.E. Ferrero, E. Magno Caldognetto and K. Vagges, *Bimodal Recognition of Italian Plosives*, Proceedings of ICPhS-95, XIII International Congress of Phonetic Sciences, Stochkolm, 14-18 August, 1995, Vol. 4, pp. 260-263.

[27] P. Cosi, M. Dugatto, F.E. Ferrero, E. Magno Caldognetto and K. Vagges, *Phonetic Recognition by Recurrent Neural Networks Working on Audio and Visual Information*, 1996 (to be published in Speech Communication Journal).

[28] P. Cosi and E. Magno Caldognetto, *Lip and Jaw Movements for Vowels and Consonants: Spatio-Temporal Characteristics and Bimodal Recognition Applications*, in Speechreading by Man and Machine: Models, Systems and Applications, NATO ASI series, 1996 (to be published by Springer-Verlag).

[29] D.W. Massaro, *Speech Perception by Ear and Eye: a Paradigm for Psychological Inquiry*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1987.

[30] B. Dodd and R. Campbell, Eds., *Hearing by Eye: The Psychology of Lip-Reading*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1987.

[31] A. MacLeod and Q. Summerfield, *Quantifying the Contribution of Vision to Speech Perception in Noise*, British Journal of Audiology, 21, 1987, pp. 131-141.

[32] G. Ferrigno and A. Pedotti, *ELITE: A Digital Dedicated Hardware System for Movement Analysis via Real-Time TV Signal Processing*, IEEE Transactions on Biomedical Engineering, BME-32, 1985, pp. 943-950.