RICONOSCIMENTO AUTOMATICO DI VOCALI PRESEGMENTATE IN PARLATO CONTINUO MEDIANTE RETI NEURALI ARTIFICIALI

Piero Cosi
Centro di Studio per le Ricerche di Fonetica C.N.R. Padova
P.zza Salvemini 13, 35131 - Padova (Italy)
Tel: 049 8755106
Fax: 049 9754560
EMail: cosi@csrf.pd.cnr.it

SOMMARIO

Viene descritto un esperimento di riconoscimento automatico di alcune vocali dell'italiano all'interno di parlato continuo. Le vocali sono state presegmentate utilizzando un sistema semiautomatico di segmentazione recentemente sviluppato su Personal Computer [1]. La classificazione degli stimoli è stata affidata ad una Rete Neurale Artificiale (RNA) addestrata al riconoscimento automatico delle sette vocali italiane [2]. Dopo una breve descrizione dei vari moduli componenti il sistema vengono illustrati e discussi i principali risultati ottenuti relativamente alla frase 'campione' registrata in occasione delle "Seconde Giornate di Studio del Gruppo di Fonetica Sperimentale (A.I.A.)" [3].

INTRODUZIONE

Tre ipotesi di lavoro sono state formulate per lo sviluppo del sistema di riconoscimento utilizzato per la realizzazione di questo esperimento:

- le trasformazioni provocate dalla coclea e dalle connesse fibre nervose sul segnale vocale danno luogo a 'pattern' di risposta neurale le cui proprietà differiscono in modo significativo dalle proprietà delle rappresentazioni spettrali fornite dalle usuali tecniche di analisi del segnale verbale (FFT, LPC, Cepstum, etc. etc.). Essenzialmente, tali tecniche non catturano tutte quelle informazioni relative ai fenomeni dinamici e non lineari che sono invece ampiamente utilizzate dal nostro apparato uditivo periferico durante il processo di riconoscimento;

- il processo interpretativo del messaggio verbale è guidato da alcune "zone di elevata affidabilità" in cui il sistema si focalizza inizialmente per poi rivolgere l'attenzione alle altre zone di più difficile interpretazione ("island driving search strategy" [4]);

- le informazioni necessarie e sufficienti per il riconoscimento del messaggio verbale sono distribuite e vengono lentamente apprese da esempi preventivamente presentati al sistema.

Sulla base della prima ipotesi di lavoro, per quanto riguarda l'analisi acustica del segnale vocale, si è utilizzato un modello del sistema uditivo periferico (MSUP) [5], mentre, in conseguenza della seconda assunzione ed in particolare assegnando alle vocali o più in generale alle sequenze vocaliche, ossia alle regioni ad elevato contenuto energetico e spettrale, il compito di rappresentare queste "isole" di sicurezza da cui partire nel processo interpretativo, si è utilizzato un algoritmo di segmentazione semiautomatica., recentemente sviluppato su Personal Computer [1], per enucleare gli stimoli vocalici da fornire in ingresso alla vera e propria struttura di riconoscimento, rappresentata da una Rete Neurale Artificiale, opportunamente addestrata al riconoscimento delle vocali italiane [2]. La scelta di una tale architettura è infatti in linea con la terza ipotesi di lavoro in quanto è stato ampiamente dimostrate come le RNA siano un paradigma computazionale molto interessante per l'apprendimento ed il riconoscimento di particolari unità o caratteristiche fonetiche [6], [7], vista la loro abilità di "apprendere" e "generalizzare" le caratteristiche discriminanti gli stimoli in ingresso esclusivamente dagli esempi che vengono loro somministrati in fase di apprendimento. Tale caratteristica le rende infatti assai interessanti se paragonate ai ben più complessi sistemi in cui ogni conoscenza deve essere esplicitata dallo sperimentatore.

METODO

In Figura 1 è illustrato un diagramma a blocchi del sistema di riconoscimento sviluppato in questo lavoro. Il segnale vocale, è stato analizzato mediante un modello fortemente ispirato alla fisiologia del sistema uditivo periferico umano sviluppato da S. Seneff [5] la cui efficacia quale tecnica di analisi è stata ampiamente dimostrata in esperimenti di segmentazione e riconoscimento automatici soprattutto in condizioni di basso rapporto segnale rumore [8].

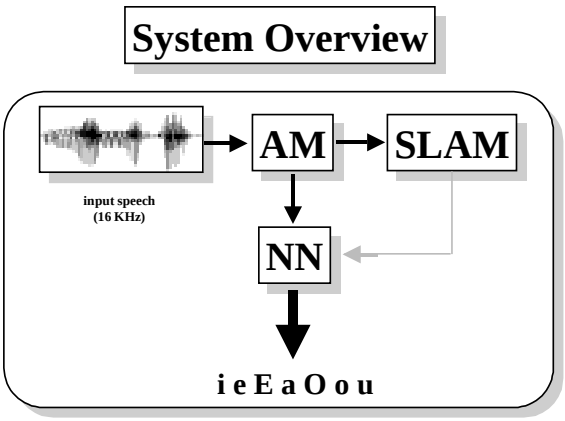

Figura 1. Diagramma a blocchi del sistema. Il segnale viene analizzato dal modello del sistema uditivo (AM) segmentato da SLAM e le zone vocaliche sono riconosciute da una Rete Neurale Artificiale addestrata al riconoscimento delle sette vocali italiane.

Per quanto riguarda l'individuazione delle sequenze vocaliche contenute all'interno del segnale continuo in ingresso, non potendo ancora disporre di un sistema completamente automatico di segmentazione in grado di fornire basse percentuali di errore, lo si è simulato con un sistema semiautomatico denominato SLAM recentemente sviluppato su Personal Computer [1] basato sulla teoria della segmentazione multi-livello [9].

La classificazione degli stimoli individuati durante la prima fase è stata affidata ad un'architettura basata su una Rete Neurale Artificiale (RNA) addestrata al riconoscimento delle sette vocali italiane [2]. Pur essendo stata addestrata in condizioni statiche, in cui le vocali in ingresso sono rappresentate da un unico vettore di parametri corrispondente a più frames per ogni stimolo, durante la fase di test la rete in esame è stata utilizzata in modo dinamico. Ad ogni 'frame' del segnale in ingresso, corrispondente ad una delle zone vocaliche semiautomaticamente enucleate in fase di segmentazione, si è quindi cercato di associare una delle sette vocali., per riconoscere quindi i singoli frames del segnale in ingresso.

ESPERIMENTO

Il sistema non è in 'real-time' ed è realizzato in parte su Personal Computer (Intel DX2 50MHz) ed in parte su DEC VAXstation 3100 collegate in rete DEC PathWorks 4.1[1], come illustrato nelle Figure 2 e 3.

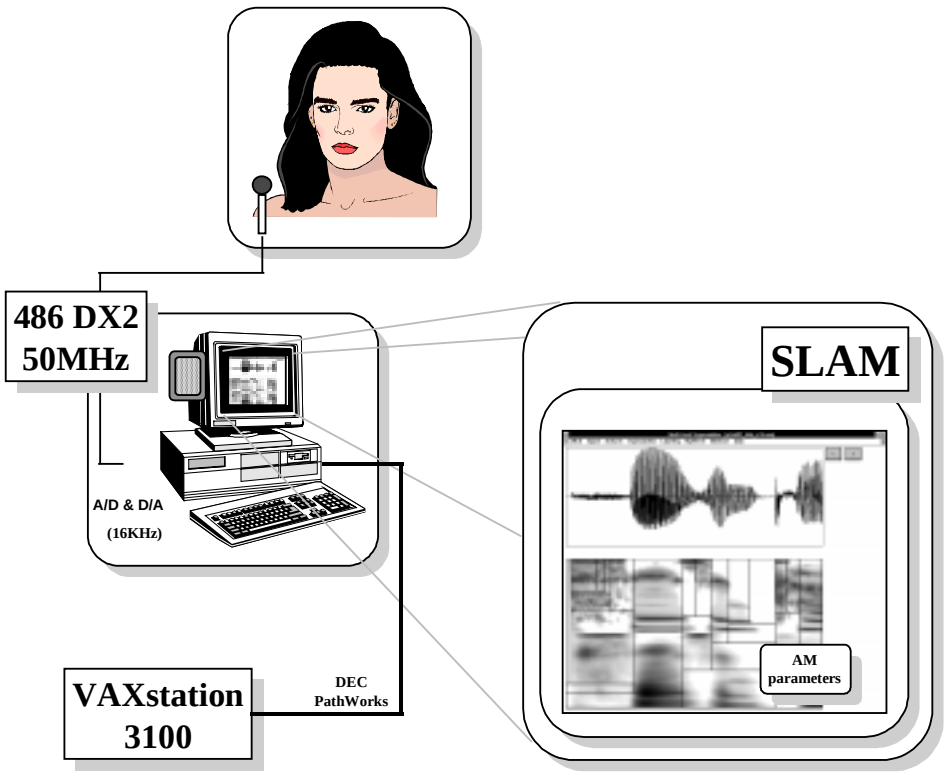

Figura 2. Acquisizione e segmentazione (SLAM) sono eseguite su Personal Computer.

[1] Intel DX2, DEC VAXstation 3100, DEC PathWorks 4.1 sono 'copyright' delle rispettive case di produzione: Intel e Digital.

Al PC sono riservati l'acquisizione del segnale e l'applicazione di SLAM per la segmentazione degli stimoli, mentre alla VAXstation 3100 sono riservati l'analisi con il MSUP e l'applicazione della RNA per la classificazione degli stimoli. Il segnale vocale, i parametri estratti dal MSUP, utilizzati sia da SLAM che dalla RNA, e le segmentazioni prodotte da SLAM sono raccolti in un database memorizzato su disco, sulla VAXstation 3100, che può essere letto direttamente dal PC mediante la rete DEC PathWorks.

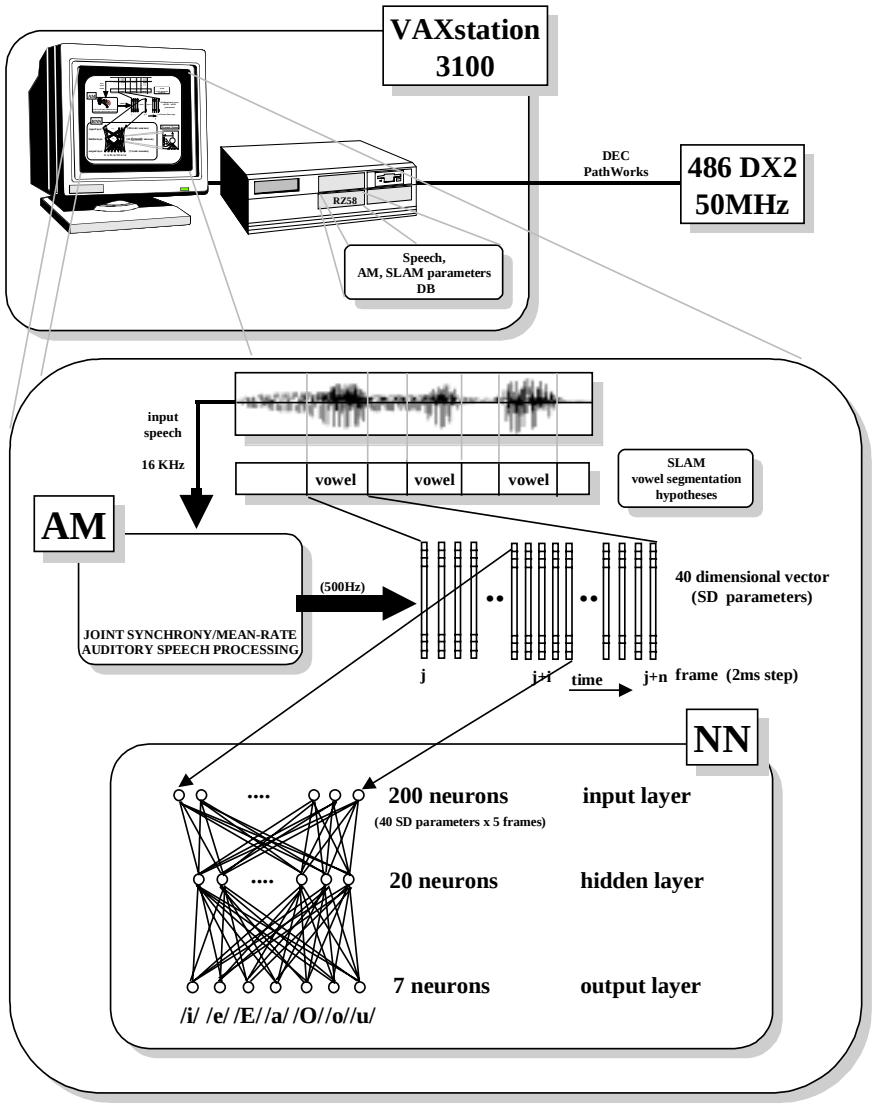

Figura 3. L'analisi con il Modello del sistema uditivo ed il riconoscimento delle zone vocaliche individuate da SLAM sono eseguite su VAXstation 3100.

Il segnale vocale utilizzato per questo esperimento è costituito dalla frase 'campione' registrata in occasione delle "Seconde Giornate di Studio del Gruppo di Fonetica Sperimentale (A.I.A.)" [3]. La frase campione è stata divisa in 16 parti più semplici allo scopo di semplificare l'esame dei risultati anche se il sistema poteva essere applicato all'intero stimolo. Queste 16 parti sono state elaborate dal MSUP e successivamente segmentate da SLAM allo scopo di enucleare le zone vocaliche.

In Figura 4 e 5 sono illustrati due esempi di segmentazione prodotti da SLAM relativamente alle frasi "...vagonate di giapponesi..." e "...gli vendono...". La RNA utilizzata per classificare le zone vocaliche all'interno dei vari stimoli è illustrata in Figura 6. Tale rete è stata addestrata al riconoscimento delle 7 vocali italiane su un database acquisito in condizioni completamente diverse ed ovviamente con parlanti differenti da quello relativo alla frase campione utilizzato in questo esperimento. E' immediato verificare in letteratura che quasi tutti i lavori sui più importanti sistemi di riconoscimento presentati in vari convegni internazionali sono invece relativi ad esperimenti effettuati su dati "uniformi", almeno per quanto riguarda le modalità di registrazione degli stimoli.

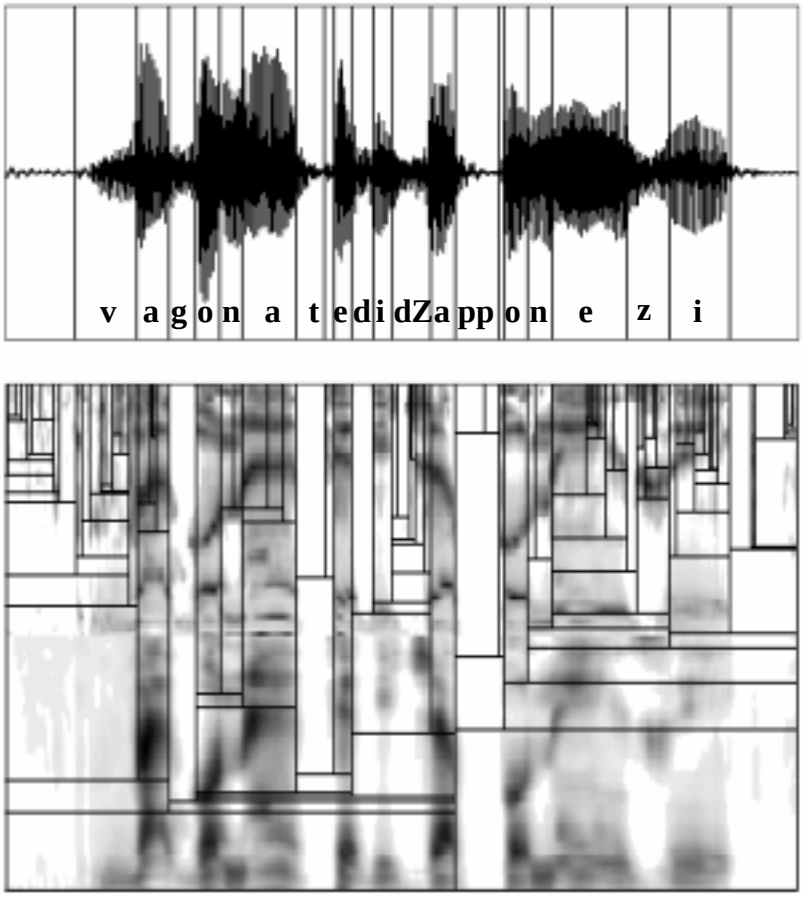

Figura 4. Segmentazione eseguita con SLAM della frase "...vagonate di giapponesi...".

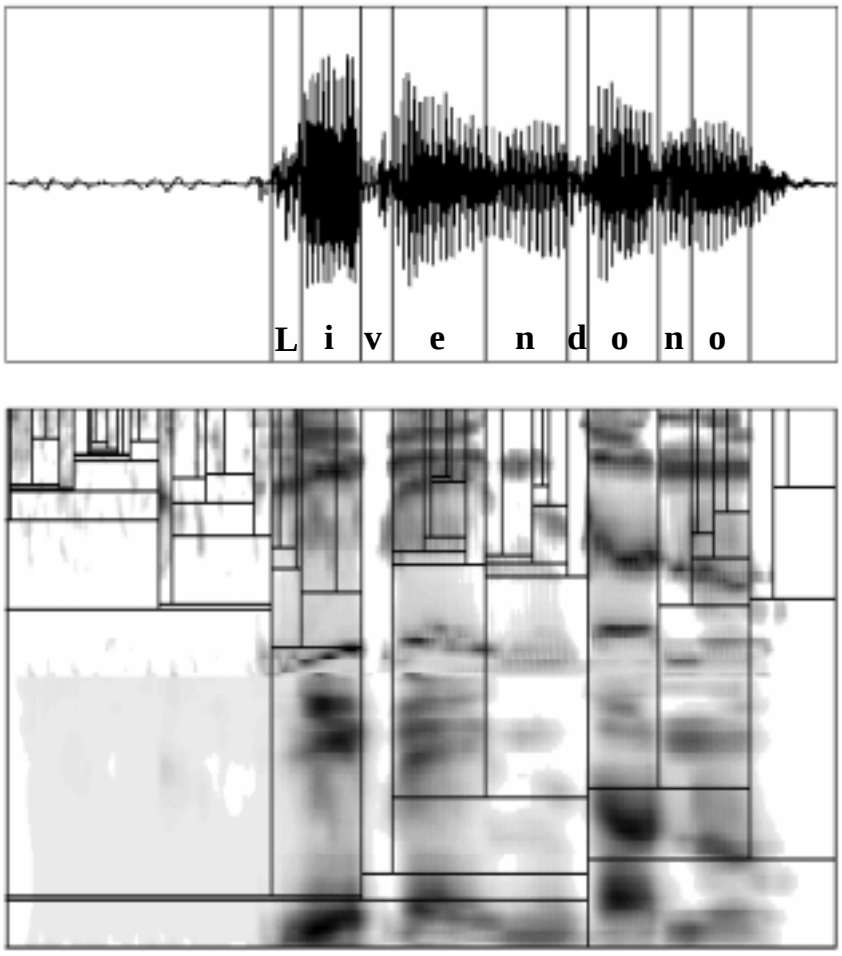

Figura 5. Segmentazione eseguita con SLAM della frase "...gli vendono...".

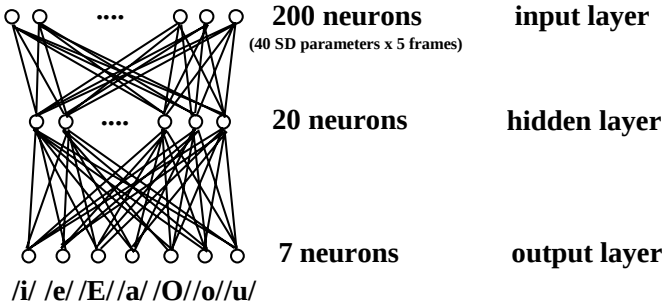

Figura 6. Struttura della rete neurale addestrata al riconoscimento delle vocali italiane.

Nelle Figure 7, 8 e 9 sono illustrate le risposte del sistema relativamente alle frasi "Che senso ha...", "scaraventare" e "le schifezze", dalle quali è possibile notare l'efficacia della classificazione. Nelle Figure la risposta della RNA è indicata, solo in corrispondenza delle zone vocaliche presegmentate con SLAM, ed in particolare vengono evidenziate le vocali riconosciute correttamente oppure quelle ad esse immediatamente adiacenti nel "triangolo delle vocali" dell'italiano. In Figura 7, in particolare, è da notare la risposta del sistema (/O/ = 'o' aperta in SAMPA [10]), in coincidenza della transizione all'interno della sequenza /oa/, mentre in figura 8 si possono osservare due apparenti errori di riconoscimento per quanto riguarda le prime due vocali /a/ della frase "scaraventare" riconosciute /e/ oppure /E/. Dopo un attento ascolto della frase in esame una tale risposta da parte del sistema può essere facilmente giustificata. In Figura 9 sono indicati alcuni errori di classificazione corrispondenti alle vocali /e/ ed /i/ della frase "le schifezze". Su tali vocali si sono concentrati gli errori di riconoscimento più frequenti.

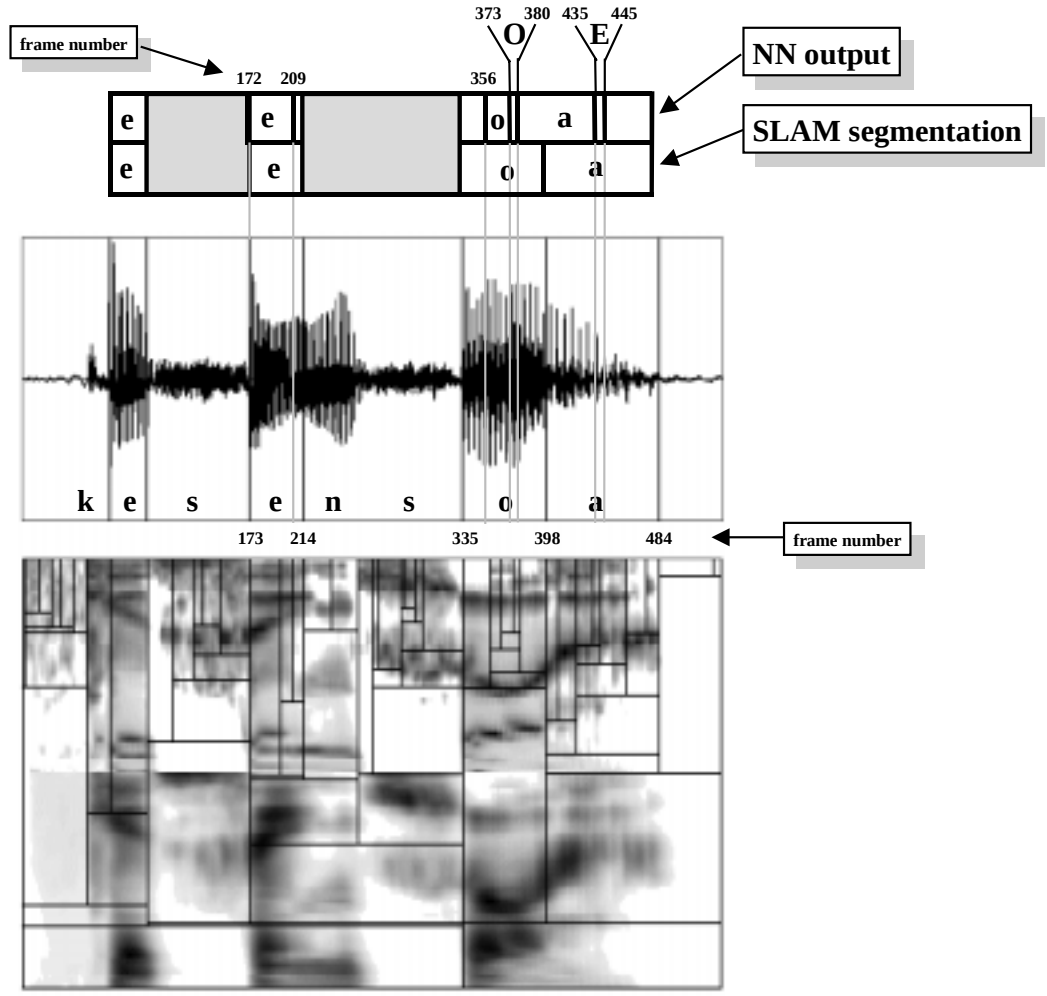

Figura 7. Riconoscimento delle zone vocaliche all'interno della frase "Che senso ha...".

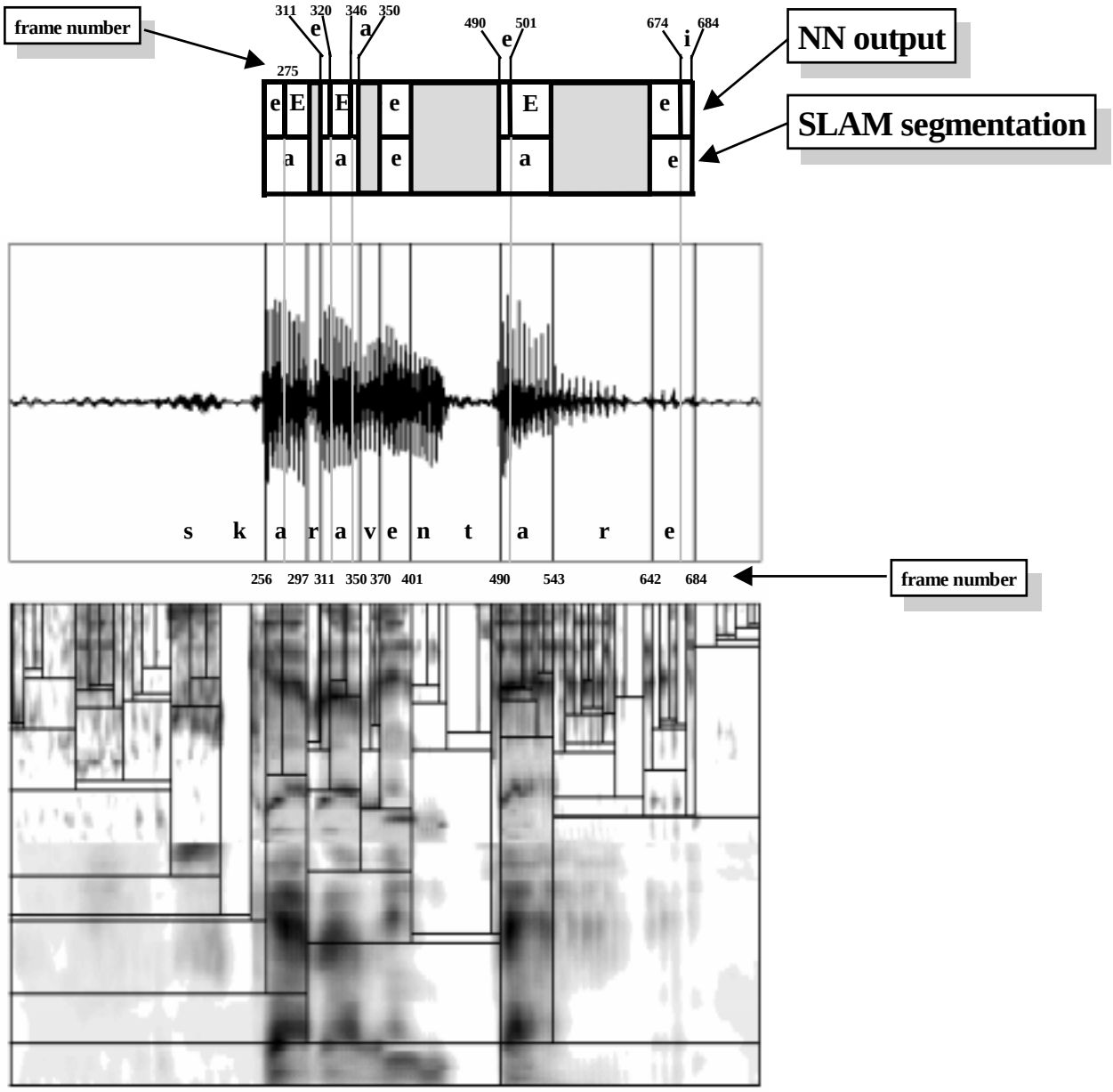

Figura 8. Riconoscimento delle zone vocaliche all'interno della frase "...scaraventare...".

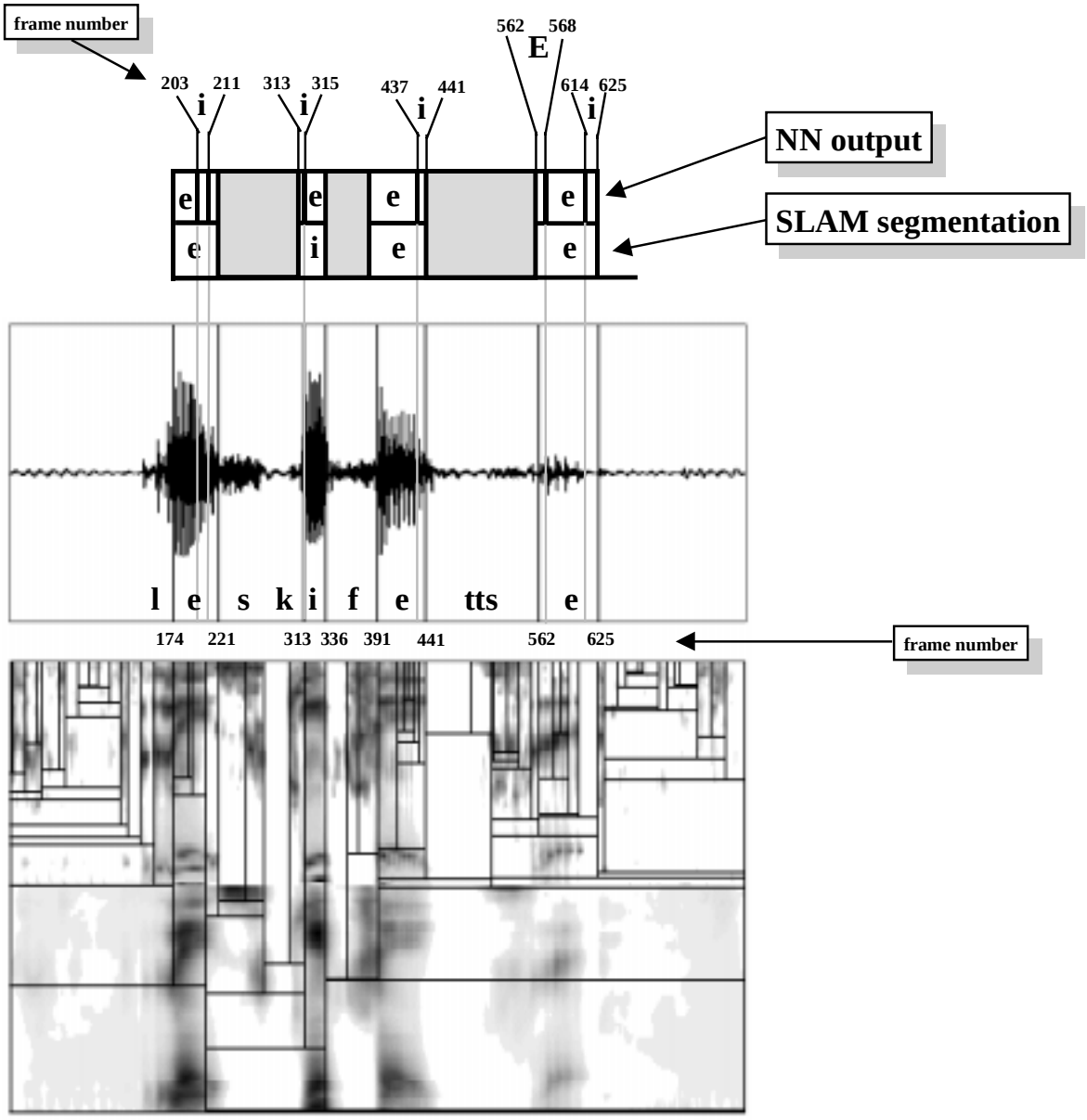

Figura 9. Riconoscimento delle zone vocaliche all'interno della frase "...le schifezze...".

CONCLUSIONI

L'utilizzazione di un modello del sistema uditivo periferico, di un sistema di segmentazione multi-livello e di una Rete Neurale Artificiale hanno consentito la realizzazione di un prototipo di un sistema semiautomatico di riconoscimento delle zone vocaliche all'interno di parlato continuo. I risultati di un semplice esperimento preliminare hanno dimostrato l'efficacia del sistema e hanno giustificato la strategia adottata.

BIBLIOGRAFIA

[1] P. Cosi (1993), "*SLAM: Segmentation and Labelling Segmentation Module*", Proceedings of EUROSPEECH-93, Berlin, 21-23 September, 1993, pp. 665-668.

[2] P. Cosi, Y. Bengio and R. De Mori (1989), "*Riconoscimento automatico delle vocali: reti neurali e un modello del sistema uditivo*", Atti XVII Convegno Nazionale AIA, Parma, 12-13 Aprile, 1989, pp. 513-518.

[3] P. Cosi (1993), "*Segmentazione semi-automatica del parlato mediante applicazione di un modello del sistema uditivo periferico*", Collana degli Atti AIA, Vol. XIX, 1993, pp. 25-38.

[4] L.D. Erman and V.R. Lesser, "*The Hearsay II Speech Understanding System: A Tutorial*", in Trends in Speech Recognition, W.A. Lea editor, Prentice-Hall, NJ, 1980, pp. 361-381.

[5] S. Seneff (1988), "*A Joint Synchrony/Mean-Rate Model of Auditory Speech Processing*", Journal of Phonetics, January 1988, pp. 55-76.

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

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

[8] P. Cosi (1992), "*Ear 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.

[9] J.R. Glass (1988), "*Finding Acoustic Regularities in Speech: Application to Phonetic Recognition*", Ph. D. Thesis, MIT Press, May 1988.

[10] A.J. Fourcin, G. Harland, W. Barry and W. Hazan eds. (1989), "*Speech Input and Output Assessment, Multilingual Methods and Standards*", Ellis Horwood Books in Information Technology, 1989.