SCUOLA
NORMALE
SUPERIORE
PISA

LO SPAZIO DELLE VOCALI

Basilio Calderone, Silvia Calamai*
Scuola Normale Superiore, Pisa; Università degli Studi di Siena – sede di Arezzo
b.calderone@sns.it; calsilvia@tiscali.it

(in corso di stampa negli Atti del III Convegno Nazionale AISV "Scienze Vocali e del Linguaggio" Metodologie di Valutazione e Risorse Linguistiche, Trento 29.IX-I.XII.2007)

Always end the name of your child with a vowel,
so that when you yell the name will carry
(Bill Cosby)

1. SOMMARIO

Il presente lavoro ha come finalità lo studio delle variabili descrittive maggiormente coinvolte in una plausibile (e automatica) differenziazione, in categorie fonemiche, di segmenti vocalici. In particolare, la nostra indagine è focalizzata sull'analisi della granularità descrittiva necessaria a rappresentare la struttura e la disposizione degli elementi vocalici all'interno di un sistema – quello pisano – che presenta alcuni aspetti tipizzanti (come l'abbassamento delle vocali toniche medio-basse e la velarizzazione di /a/). Cercheremo di scoprire come sia possibile evidenziare analogie e similarità all'interno di un *set* di vocali adottando una prospettiva *non-supervisionata*, in cui cioè non vi sia presenza di un supervisore esterno che informi il sistema sulle corrette modalità della ristrutturazione del dato in categorie: al contrario, il processo di organizzazione delle vocali, risultante a fine apprendimento, avviene meramente sulla base di regolarità e ricorrenze statistiche del *corpus* di *input*.

Per le simulazioni è stato impiegato un modello neurale ad apprendimento competitivo e non supervisionato, la cosiddetta SOM (*Self-Organizing Maps*)[1]. Quattro *corpora* vocalici, suddivisi per tipologia accentuale (presenza/assenza di accento) e stile (letto *vs.* spontaneo) sono stati scelti come *traning set*.

Il nostro inventario vocalico è definito da differenti tipologie di variabili: valori spettrali, differenze spettrali, frequenza fondamentale, durata. Al fine di verificare l'economia descrittiva necessaria a garantire una coerente e categorialmente ben definita organizzazione (*clustering*) spaziale delle vocali, del tutto automatica *e non-supervisionata*, differenti addestramenti sono stati condotti variando il numero e la tipologia delle variabili prese in esame per la rappresentazione del dato acustico. Tale procedura sperimentale è tesa a scoprire la descrizione minimale del dato, ovvero quali e quante variabili descrittive sono sufficienti a garantire un'organizzazione dei dati vocalici in *clusters* distintivi.

2. CATEGORIE, GRANULARITÀ DESCRITTIVA, AUTO-ORGANIZZAZIONE

Con il termine 'auto-organizzazione' si intende la capacità esibita dal sistema di raggiungere, nel tempo, uno stato di equilibrio o, più in generale, una ri-strutturazione

* Il lavoro è frutto della collaborazione dei due autori. Ai fini concorsuali, sono da attribuirsi a BC i paragrafi 2, 3, 5.2; a SC i paragrafi 1, 4, 5.1. Il paragrafo finale è opera di entrambi.

[1] Kohonen (2001).

ultima delle unità componenti il sistema stesso, in accordo a principi di similarità strutturale e regolarità statistiche (Ref). Nel nostro caso il sistema coincide con l'architettura di una SOM addestrata mediante dati acustici di segmenti vocalici.

Una caratteristica fondamentale di questi processi di strutturazione dinamica è che essi non rispondono a regole o criteri di ottimizzazione e strutturazione globali, ma obbediscono esclusivamente a principi di interazione locale i cui effetti si ripercuotono tuttavia a livello dell'intera struttura dei dati. In quest'ottica, appare evidente come la rappresentazione dell'*input* abbia un ruolo cruciale per un possibile processo di strutturazione (coerente) del dato e per favorire così l'emergenza di *clustering* categoriali. La SOM perviene, in maniera del tutto non supervisionata, a una organizzazione spaziale (solitamente su uno spazio bi-dimensionale) dei dati di ingresso sulla base di un qualche criterio di similarità: dati simili saranno disposti spazialmente vicini all'interno della mappa. In questo modo, la mappa evidenzia delle categorie (spaziali) del dato (le vocali, nel nostro caso) che verranno organizzate sulla base di attributi qualitativi condivisi da tutti gli elementi di una medesima categoria e riconosciuti, al contempo, dal sistema.

Una SOM raggiunge quindi, ad apprendimento ultimato, un livello di *granularità descrittiva* dei dati di ingresso che si traduce in una categorizzazione automatica di quest'ultimi sulla base della natura e della similarità dei dati. Tale *granularità descrittiva*, ovviamente, dipenderà da diversi fattori tra cui il numero dei neuroni definenti una SOM (maggiore sarà il numero e potenzialmente più segmentata e capillare sarà l'organizzazione in categorie), la frequenza (dati più volte reiterati all'interno del *training set* tenderanno a formare una propria categoria spaziale, non condivisa con altri), la rappresentazione del dato stesso (numero e tipologia di variabili considerate nella rappresentazione del dato).

Come mostreremo tra breve, la nostra analisi verterà su una possibile modellizzazione della *granularità descrittiva,* esibita dalla SOM, di dati vocali rispetto al *set* di variabili considerati per la loro rappresentazione. La modellizzazione lascerà quindi inalterato, nel nostro ciclo di simulazioni, il numero di neuroni definenti la SOM e gli indici di frequenza dei dati nei nostri *corpora*, e modificherà invece solo il numero di variabili usate nella descrizione del dato. Si cercherà, in altri termini, di definire un'economia descrittiva del dato, capace di garantire una coerente discriminazione in categorie vocaliche con la minor rappresentazione possibile.

Il paragrafo seguente presenta in dettaglio le caratteristiche basiche di una SOM: una panoramica della sue caratteristiche è in 3.1; informazioni di carattere più tecnico riguardanti l'algoritmo di apprendimento sono in 3.2; gli aspetti e le proprietà della SOM, utili per la nostra indagine, vengono enfatizzati in 3.3.

3. LA SOM (*SELF-ORGANIZING MAP*)

3.1 Caratteristiche e funzionamento

Con SOM (*Self-Organizing Map*) si intende una struttura connessionistica i cui neuroni sono organizzati all'interno di un *array* di superficie mono-dimensionale o più spesso bi-dimensionale (Figura 1). La mappa riceve in ingresso dei vettori di *input* e, ad addestramento ultimato, è in grado di esibire sulla sua superficie una compressione topologica di tutti i dati vettoriali acquisiti. In tal modo similarità e analogie (spaziali) dei vettori presentati in ingresso riceveranno medesimo trattamento dalla SOM che provvederà a ordinarli su intorni spaziali prossimi. A una prima analisi, la SOM svolge il ruolo di *clusterizzatore* di dati d'ingresso, i quali vengono accomunati spazialmente in base a una qualche loro caratteristica più o meno soggiacente. Nelle migliori delle ipotesi, ad ogni

neurone della rete corrisponderebbe un *cluster* e cioè un raggruppamento omogeneo dei dati di *input*, localizzati in quella zona per un qualche criterio di similarità individuato dalla rete[2].

Il processo di addestramento (*training process*) coincide quindi con una fase di adattamento spaziale agli *input* di natura vettoriale. Tale adattamento è compiuto dalla SOM tramite modifica dei propri coefficienti di connessione.

Come già rilevato, il processo di apprendimento è non-supervisionato, in altre parole non vi è definizione in anticipo sulla configurazione spaziale cui perverrà la SOM. La variazione dei coefficienti di connessione, e quindi la configurazione della mappa, dipenderà soprattutto dalle caratteristiche statistiche del campione di dati che costituisce la base di conoscenza su cui si informerà tutto il processo di apprendimento.

Ogni qualvolta si presenta un *pattern* di ingresso, viene infatti eseguito un *matching* tra quel *pattern* (definito come un vettore a *n*-dimensioni) e tutte le unità neuronali formanti la mappa bi-dimensionale (definite anch'esse in termini vettoriali a *n*-dimensioni). Il criterio adottato per stimare la diversità tra i due vettori fa capo alla nozione di 'distanza'. La 'distanza' in questione può essere calcolata in vari modi (solitamente si considera la distanza euclidea). È importante notare che questo calcolo è reso possibile dal fatto che i due vettori, quello che rappresenta il *pattern* in ingresso e quello che rappresenta l'unità della mappa, contengono il medesimo numero di componenti.

Si considera ora come unità di uscita 'vincitrice' quella in cui la 'distanza' precedentemente calcolata assume il minimo valore. L'algoritmo di Kohonen adotta la strategia della 'bolla': non vi è una sola unità 'vincitrice', ma si considerano 'vincitrici' (ovvero si recuperano) anche tutte le altre unità di uscita che si trovano entro una certo 'intorno' dall'unità 'vincitrice' principale. Tale 'intorno', nella terminologia di Kohonen, è detto *raggio della bolla*.

Dopo aver individuato le unità 'vincitrici', si modificano solo i coefficienti di connessione delle linee che fanno capo a queste unità, in modo da diminuire la 'distanza' tra loro e il segnale presentato in ingresso. Si passa poi al vettore di *input* successivo. Man mano che il processo di apprendimento va avanti, il 'raggio' della bolla e il tasso di modifica dei coefficienti vengono fatti progressivamente decrescere, con una legge definita dallo sperimentatore.

[2] Non è esclusa però la possibilità, invero più frequente, che sia un assemblaggio di neuroni a definire un unico *cluster*. Ciò è dovuto alla risoluzione della mappa, che è in relazione con la sua dimensione: aumentando, infatti, i nodi (e quindi anche la dimensione), questi individueranno una categoria più ristretta e più granulare.

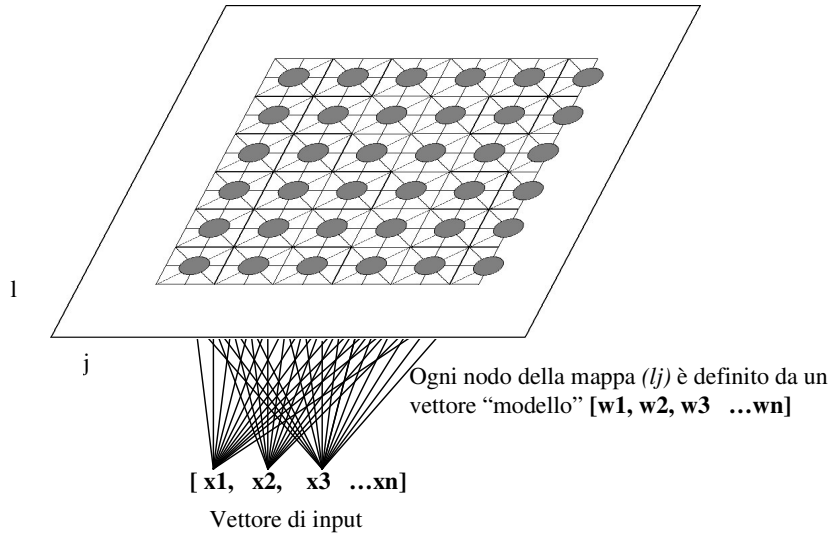

Figura 1: SOM 6 x 6.

Quando quest'ultimo decide di interrompere il processo di apprendimento, se i *patterns* di ingresso presentano al proprio interno qualche regolarità statistica, si forma spontaneamente una differenziazione nei ruoli delle unità di uscita. Ognuna di queste, infatti, risponde (cioè diventa 'vincitrice') solo in corrispondenza a particolari classi di *pattern* di ingresso.

3.2 L'algoritmo di addestramento

Il *set* degli esempi di *input* è descritto da un vettore $\mathbf{x}(t) \in R^n$ dove t indica una discretizzazione temporale. Ogni nodo i della rete SOM contiene un vettore 'modello' $\mathbf{m}_i(t) \in R^n$, che ha lo stesso numero di elementi del vettore di input $\mathbf{x}(t)$.

L'algoritmo della SOM raggiunge uno stato di equilibrio, creando quindi degli ordini spaziali (*clusters*), tramite la ripetizione dei seguenti passaggi:

1. Ogni vettore di input $\mathbf{x}(t)$ è confrontato con tutti i vettori 'modello' $\mathbf{m}_i(t)$ al fine di trovare la BMU (*best-matching unit*), l'unità cioè che presenta maggiori similarità (in termini di distanza euclidea) con il vettore di *input*. Il BMU è spesso chiamata l'unità 'vincitrice' e si indica con **c**.
2. Ogni vettore 'modello' delle unità 'vincitrici' e un numero di unità circostanti **c** (definenti la già conosciuta nozione di 'bolla') sono adattati[3] al vettore di ingresso secondo una regola di apprendimento che esponiamo di seguito.

L'idea fondamentale nel *SOM learning process* è che per ogni vettore di *input* $\mathbf{x}(t)$ (ordinato in un apposito *training set*), il raggio della 'bolla' decresca gradualmente (fase di

[3] Nei termini dei propri coefficienti di connessione, si intende.

apprendimento) e l'unità 'vincitrice' selezionata subisca una modifica (vettoriale) al fine di adattarsi il più possibile a $\mathbf{x}(t)$, il vettore spaziale di ingresso.

Durante il processo di apprendimento possono verificarsi contraddizioni individuali, ma nel complesso il risultato della rete è l'emergere di valori ordinati per $\mathbf{m}_i(t)$ localizzati sulla mappa.

Se il numero degli esempi di *input* disponibili è basso, essi devono essere presentati alla rete in maniera re-iterativa.

La definizione di **c** (l'unità 'vincitrice') è data dalla differenza della distanza euclidea tra il vettore d'ingresso ed il vettore 'modello', ottenuta grazie all'espressione:

$$\mathbf{c} : \mathbf{m}_c(t) = \mathbf{min}\left\|\mathbf{x}(t) - \mathbf{m}_i(t)\right\| \qquad (1)$$

L'adattamento è implementato come una graduale riduzione della differenza tra i componenti del vettore di input e i componenti del vettore 'modello'. L'algoritmo è espresso nel seguente modo:

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + \alpha(t)\left[\mathbf{x}(t) - \mathbf{m}_i(t)\right] \quad \text{per ogni } i \in N_c(t) \quad (2)$$

dove t rappresenta la discretizzazione temporale, il parametro $\alpha(t) \in [0, 1]$ è uno scalare, detto anche *indice di apprendimento* e decresce durante le fasi dell'apprendimento. La grandezza della 'bolla', per il recupero e l'aggiornamento delle unità circostanti **c**, è definito da $N_c(t)$ e anch'esso decresce durante l'adattamento.

Uno degli aspetti più interessanti della regola di apprendimento è che essa contiene due parametri variabili col tempo: la grandezza definente la 'bolla', N_c, e l'indice di apprendimento, α. Generalmente essi vengono fatti diminuire col tempo anche se in alcune applicazioni α viene lasciato costante. Solitamente la grandezza della 'bolla' è sintetizzata dal suo raggio $r(t)$. Esempi di leggi di variazione del raggio della 'bolla' con il tempo sono

$$r(t) = \mathbf{A} - \mathbf{B} \cdot t$$

in cui il valore tipico di A è 4 e di B è 0.0002, oppure

$$r(t) = \mathbf{A} / \left(1 + \mathbf{B} \cdot t\right)$$

in cui i valori tipici sono A = 4, B = 0.003.

Leggi esattamente analoghe a queste possono essere utilizzate per far diminuire col tempo l'indice $\alpha(t)$. In sostanza esse servono a far sì che, nella fase iniziale dell'apprendimento, vengano interessati dal cambiamento molti coefficienti di connessione, i quali subiscono variazioni abbastanza rilevanti (valori di α e $r(t)$ elevati), mentre, al termine del processo di apprendimento, quando ormai ogni unità si è specializzata nel rispondere a particolari *pattern* in ingresso, conviene che sia il raggio della 'bolla' che l'indice di apprendimento siano molto modesti.

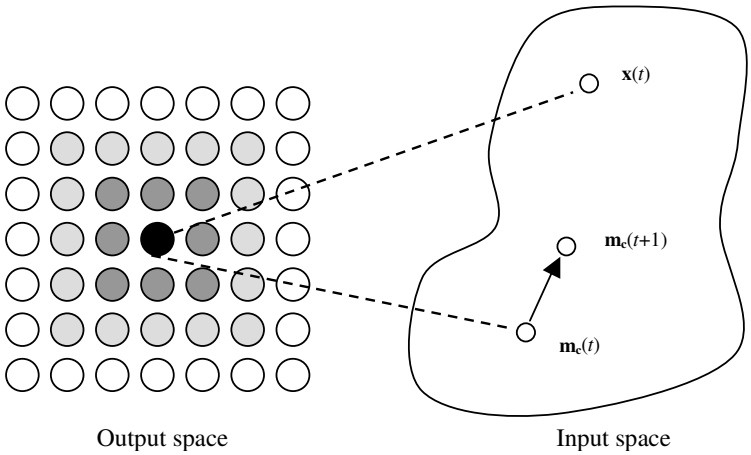

Figura 2: Istantanea della fase di adattamento tra il vettore d'ingresso, x(t), e quello della mappa $m_c(t)$.

La Figura 2 mostra una rappresentazione grafica di una *Self-Organizing Map* durante il processo di adattamento. Il cerchietto nero indica che l'unità è stata selezionata come unità 'vincitrice' per la rappresentazione del *pattern* di *input* $\mathbf{x}(t)$. Il vettore dell'unità vincente, $\mathbf{m}_c(t)$, è adattato (durante la fase di apprendimento) al *pattern* di *input* e quindi $\mathbf{m}_c(t + 1)$ sarà più vicino al vettore *input* $\mathbf{x}(t)$ di quanto non lo fosse $\mathbf{m}_c(t)$. Tale processo di adattamento, è noto, coinvolge anche un numero di unità prossime all'unità 'vincitrice'. In Figura 2 queste unità sono rappresentate mediante una scala di tonalità di grigio. Maggiore è l'intensità di grigio, maggiormente queste unità saranno coinvolte nel processo di adattamento (apprendimento).

3.3 Aspetti multipli di una SOM

La SOM può essere ritenuta un *modello di apprendimento non-supervisionato*. Come tale risponde bene alla risoluzione di problemi quali i seguenti:

- ***Clustering***. Fornito un insieme di elementi in ingresso, questi devono poter essere raggruppati in *clusters* (gruppi omogenei) e il compito della rete, dopo l'apprendimento, è individuare, per ogni ingresso, il gruppo di appartenenza.
- ***Quantizzazione vettoriale***. Il valore di ingresso fornito alla rete è uno spazio continuo che deve essere discretizzato. Gli ingressi della rete sono vettori n-dimensionali **x**, l'uscita è una rappresentazione discreta dello spazio d'ingresso. La rete deve fornire una rappresentazione ottimale della discretizzazione dello spazio d'ingresso.
- ***Estrazioni delle caratteristiche***. La SOM estrae delle caratteristiche dall'ingresso iniziale, questo spesso implica una riduzione della dimensionalità.

La SOM inoltre è usata spesso utilizzata come *dispositivo per analisi statistiche*. Una rete SOM, infatti, costituisce un metodo di riduzione e proiezione delle dimensionalità, capace di mappare (e categorizzare) *high-dimensional data space* in *low-dimensional data space*, riuscendo a preservare però le proprietà dinamiche e topologiche dello spazio iniziale.

4. LA BASE DEI DATI, LE VARIABILI, LA LETTERATURA

Il materiale su cui si basa l'analisi proviene da *corpora* differenti, diversificati per accento (vocali toniche *vs.* vocali atone) e per stile (parlato letto *vs.* parlato semispontaneo). La parte numericamente più consistente è rappresentata dal *corpus* di parlato letto dell'*Archivio delle Varietà di Italiano Parlato* (AVIP), costituito da liste di parole e pseudo-parole lette dai soggetti che hanno poi partecipato alle sessioni di *Map Task*[4]. I soggetti – sei studenti universitari, nati e vissuti a Pisa, con un'età compresa trai 23 e i 31 anni – rappresentano un campione omogeneo per sesso, età, provenienza geografica, livello socio-culturale. In totale, le vocali toniche prese in considerazione sono 1199, quelle atone 1509.

Il sistema vocalico oggetto dell'indagine è fonologicamente coincidente con quello dell'italiano *tout court*. Tuttavia, nel caso delle vocali toniche, mostra alcune specificità di carattere essenzialmente fonetico che lo rendono particolarmente 'riconoscibile' e 'identificabile', soprattutto in un'ottica sociolinguistica: in particolare, il sistema è caratterizzato da un marcato abbassamento delle vocali medio-basse (soprattutto quella anteriore) e dalla posteriorizzazione di /a/ (Calamai 2004).

I risultati delle simulazioni con il parlato letto sono stati poi confrontati con un *corpus* decisamente più ristretto – e riferito a due soli locutori – rappresentato dal parlato semispontaneo (dialoghi di tipo *map task*)[5], sia tonico (362 entrate vocaliche) che atono (147 entrate vocaliche). Il senso di un simile confronto diafasico ha un valore meramente esplorativo e più che fornire risultati intende tracciare la strada per future linee di ricerca. Del resto, vista la difficoltà – in più luoghi e da più autori ribadita – di misurare i valori formantici del parlato meno controllato non desterà stupore la pochezza numerica del campione: a quanto ci consta sono pochissimi i lavori che compiono simulazioni con stili di eloquio che risultano essere pesantemente condizionati da coarticolazione e ipoarticolazione.

L'analisi acustica è stata compiuta con il software *Multispeech*, per l'analisi formantica sono state utilizzate delle macro sviluppate da Cioni (2001)[6]. Per ogni vocale tonica e atona sono stati misurati i valori in Hertz della prima, della seconda e della terza formante e i valori della frequenza fondamentale, rilevati in tre punti del segmento vocalico (nella parte ritenuta più stabile); per ciascuna variabile è stata fatta la media aritmetica delle tre misurazioni. I dati sottoposti ad analisi sono dunque espressi mediante valori acustici; non è stata al momento compiuta alcuna normalizzazione uditiva o logaritmica.

Per le simulazioni, sono state prese in considerazione variabili pure – rappresentate dalle formanti F_1, F_2 e F_3 – e differenze spettrali, rappresentate da (F_1-f_0), (F_2-F_1) e (F_3-F_2). Nel costruire le simulazioni le variabili sono state combinate in vario modo, prendendo in considerazione sia procedure di normalizzazione cosiddette *formant-intrinsic*, sia procedure cosiddette *formant-extrinsic*, le quali oltre ai valori formantici ricorrono all'apporto della frequenza fondamentale e delle differenze spettrali.

L'inserimento della frequenza fondamentale in una indagine incentrata sui migliori criteri di rappresentazione vocalica è motivata dal fatto che questa sembra essere in grado

[4] Bertinetto (2001).

[5] Sono analizzati l'*instruction follower* della mappa A03 (Pisa, A.S.) e l'*instruction giver* della mappa B03 (Pisa, F.V.).

[6] Questi i parametri: per l'acquisizione del segnale, *sampling rate*: 11.025; per lo spettrogramma, *analysis size*: 100 punti; *window*: *Hamming*; *pre-emphasis*: 0.800; per l'analisi in LPC, *frame lenght*: 20 ms; ordine del filtro: 14; *pre-emphasis*: 0.900; per l'analisi di f0, *analysis range*: 70-350.

di agire come fattore di normalizzazione, influenzando la percezione del timbro vocalico[7]. Anche se nella teoria acustica di produzione dei suoni il filtro è indipendente dalla sorgente del segnale, è indubbio che esista una covariazione positiva tra altezza vocalica e frequenza fondamentale (il cosiddetto *intrinsic pitch*)[8].

L'utilizzo delle differenze spettrali nella rappresentazione vocalica trova la sua giustificazione negli studi di carattere percettivo condotti con stimoli sintetici. Gli esperimenti percettivi di Delattre *et alii* (1952) e poi di Carlson *et alii* (1970) hanno mostrato che la maggior parte delle vocali può essere simulata mediante oggetti sonori costituiti da due sole formanti, in cui una delle due sia rappresentata dalla media pesata di due o più formanti nelle vocali naturali. Quando le formanti sono molto vicine in frequenza (F_1 e F_2 per le vocali posteriori, F_2 e F_3 per quelle anteriori), esse vengono integrate da un punto di vista percettivo.

In letteratura sono state utilizzate numerose procedure statistiche per valutare quali siano le variabili e/o le unità di misura che più efficacemente permettono di caratterizzare i sistemi vocalici. La procedura più diffusa prevede il ricorso all'analisi discriminante (lineare o quadratica)[9], sono più scarsi – e non solo per l'analisi dei sistemi vocalici – i contributi che fanno uso di algoritmi di classificazione come CA&T[10]. Per quanto concerne il ricorso alle reti neurali, ci risulta che le SOMs non siano mai state applicate ai sistemi vocalici: le reti utilizzate per il riconoscimento e la categorizzazione sono soprattutto quelle cosiddette *feedforward* o *back-propagation* (e.g. Cosi 1991). Come si rileva anche in Jassem, Grygiel (2004: 38) c'è un certo ritardo nell'utilizzare le potenzialità delle reti nella soluzione di questioni strettamente fonetiche: questa indagine è pertanto un primo tentativo in tal senso.

5. RISULTATI

5.1 Lo spazio dei suoni: riflessi articolatori dell'organizzazione spaziale

Per le nostre simulazioni abbiamo adottato un protocollo di addestramento a due fasi:

1) una prima fase, cosiddetta *rough training phase*, in cui le dimensioni del raggio della bolla *r* e il parametro di apprendimento α hanno valori elevati e rimangono fissi durante questa prima fase di apprendimento;

2) una seconda fase, *finetuning phase*, ove *r* e α diminuiscono gradualmente.

[7] Gli studi al riguardo sono molto ampi; ci limitiamo a segnalare Traunmüller (1981); Di Benedetto (1991; 1994); Hirahara & Kato (1992), rinviando alla bibliografia ivi contenuta.

[8] Restano tuttavia ancora da chiarire le modalità con cui questa influenza di f_0 si manifesta. Keith Johnson, ad esempio, sostiene che f_0 eserciti una influenza indiretta, in altre parole questa variabile ha valore non tanto perché è inserita nella differenza spettrale (F_1-f_0) ma piuttosto perché determina l'identità del parlante: la presenza di una correlazione fra tratto vocale e frequenza fondamentale permetterebbe di descrivere l'apporto di quest'ultima alla percezione vocalica come un apporto relativo alla percezione della grandezza del tratto vocale del parlante (si tratta del cosiddetto *adjustment-to-talker model of vowel normalization*: Johnson, 1990).

[9] Sarebbe lungo ripercorrere tutti i lavori: una rassegna è in Calamai, Agonigi, Ricci (2005).

[10] Un esempio, applicato su una parte dei medesimi dati qui analizzati è Calamai, Agonigi, Ricci (2005).

Ogni addestramento è stato condotto utilizzando una SOM di dimensione 20x10 inizializzata con pesi *random* e utilizzando per la ricerca del BMU (*best-matching unit*) la distanza euclidea.

Le figure 3 e 4 riportano le organizzazioni per i quattro sistemi vocalici (tonico parlato letto, tonico parlato semispontaneo, atono parlato letto, atono parlato semispontaneo).

Ad addestramento concluso, ogni neurone della mappa diventa rappresentativo di ognuna delle unità vocaliche presenti nel *corpus* (ovviamente un neurone può rappresentare al contempo diverse vocali appartenenti anche a categorie differenti. Le Figure 3 e 4 mostrano, per agilità di visualizzazione, solo le vocali che con maggiore frequenza 'cadono' all'interno dei nodi/neuroni della SOM). Ecco una prima analisi dei risultati:

- vocali simili da un punto di vista articolatorio occupano spazi contigui nelle SOMs: in altre parole, la disposizione spaziale, letta dall'alto in basso, ricalca il *continuum* articolatorio anteriore-centrale-posteriore;
- stili iperarticolati si organizzano meglio di stili ipoarticolati: la sovrapposizione nello spazio acustico di vocali ipoarticolate trova riscontro in una rappresentazione meno limpida sul piano spaziale (il parlato semispontaneo si organizza peggio rispetto al parlato letto, e – all'interno del parlato semispontaneo – il vocalismo atono si organizza peggio del sistema tonico);
- la riconoscibilità di /ɛ/, marca diatopica per eccellenza del sistema tonico pisano, trova un riflesso nel suo isolamento spaziale;
- la posteriorizzazione di /a/ – altra caratteristica diatopica del sistema tonico pisano – è dimostrata dalla maggiore vicinanza spaziale alla classe delle vocali posteriori.

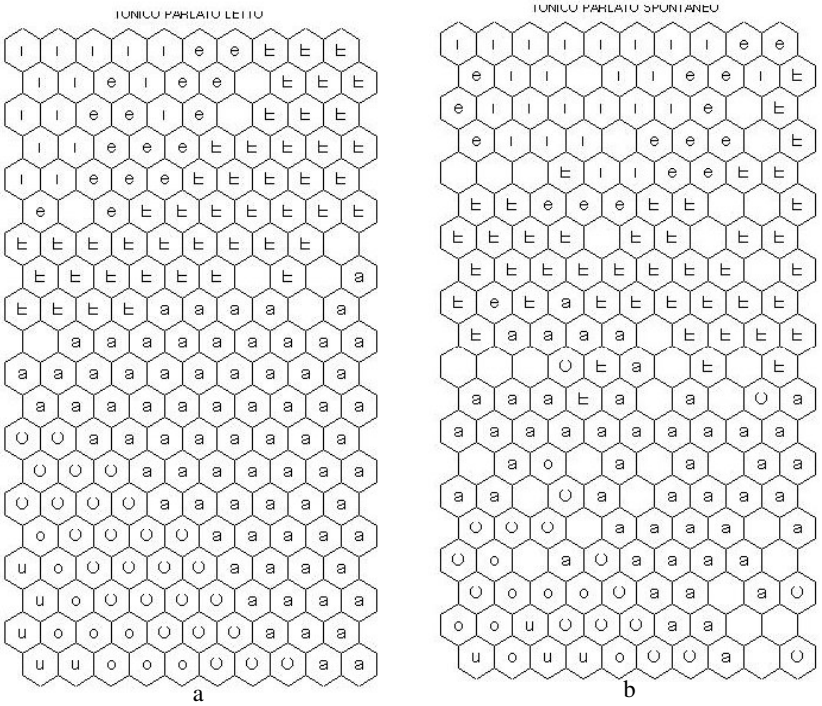

Figura 3: Organizzazione del sistema TONICO PARLATO LETTO (a) e del SISTEMA TONICO PARLATO SEMISPONTANEO (b). Per l'addestramento è stata impiegata una SOM 20X10.

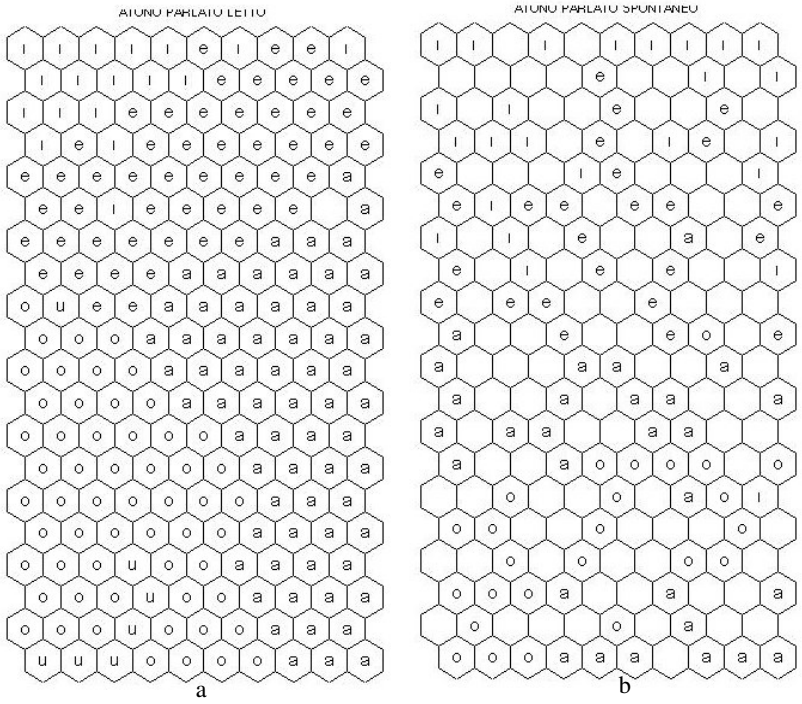

Figura 4. Organizzazione del sistema ATONO PARLATO LETTO (a) e del sistema ATONO PARLATO SEMISPONTANEO (b). Per l'addestramento è stata impiegata una SOM 20 X 10.

5.2. Economia descrittiva (less is more)

Le organizzazioni finali, esibite dalla SOM per i quattro *corpora* vocalici, hanno permesso l'individuazione di omogenee e, il più delle volte, ben distinte categorie vocaliche (come si è appena visto, la bontà delle organizzazioni varia in accordo alla differenza di stili, per cui stili iperarticolati 'funzionano meglio' di stili ipoarticolati). Una successiva analisi considererà quali variabili, usate nella rappresentazione delle vocali, appaiono essere più pertinenti per l'(auto)organizzazione del dato stesso nelle varie classi vocaliche.

Per motivi di spazio, ci limitiamo a riportare i risultati solo del TONICO PARLATO LETTO, estendendo le conclusioni della nostra indagine anche ai restanti *corpora*.

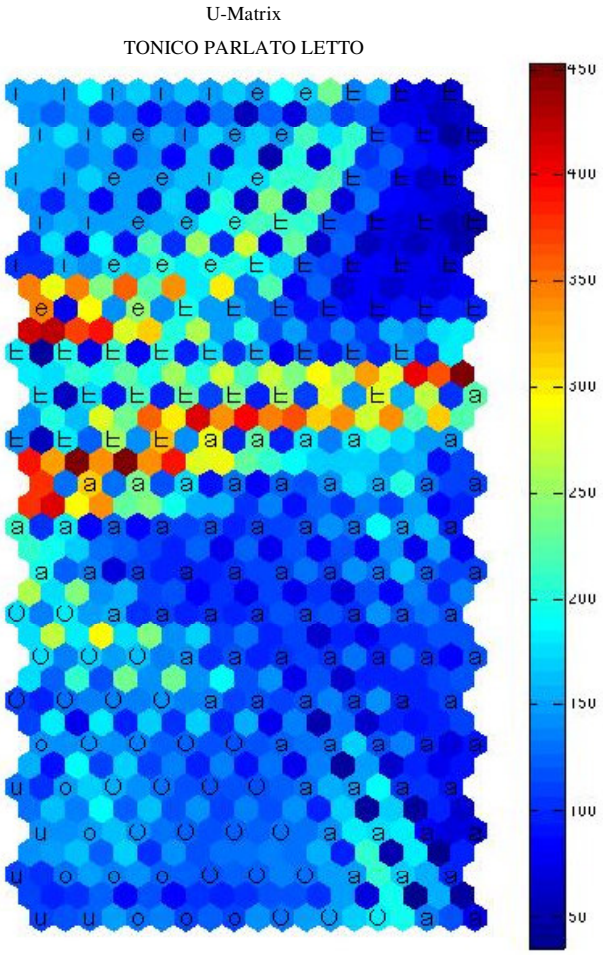

Figura 5. Visualizzazione U-matrix della SOM addestrata tramite il *corpus* TONICO PARLATO LETTO utilizzando nove varibili per la rappresentazione del *set* vocalico. Dati vicini corrispondono a valori bassi di intensità (tonalità di blu) e i *clusters* identificati dalla mappa sono separati dalle zone ad intensità maggiore (tonalità di rosso)

La possibilità di evidenziare delle classi emergenti, in maniera del tutto non-supervisionata, costituisce forse l'aspetto più interessante delle SOMs. Un metodo efficace per l'individuazione di possibili *clusters* di dati all'interno di una SOM addestrata è *la unified distance matrix* o, come si trova in letteratura, U-matrix[11].

[11] Ultsch & Siemon (1990).

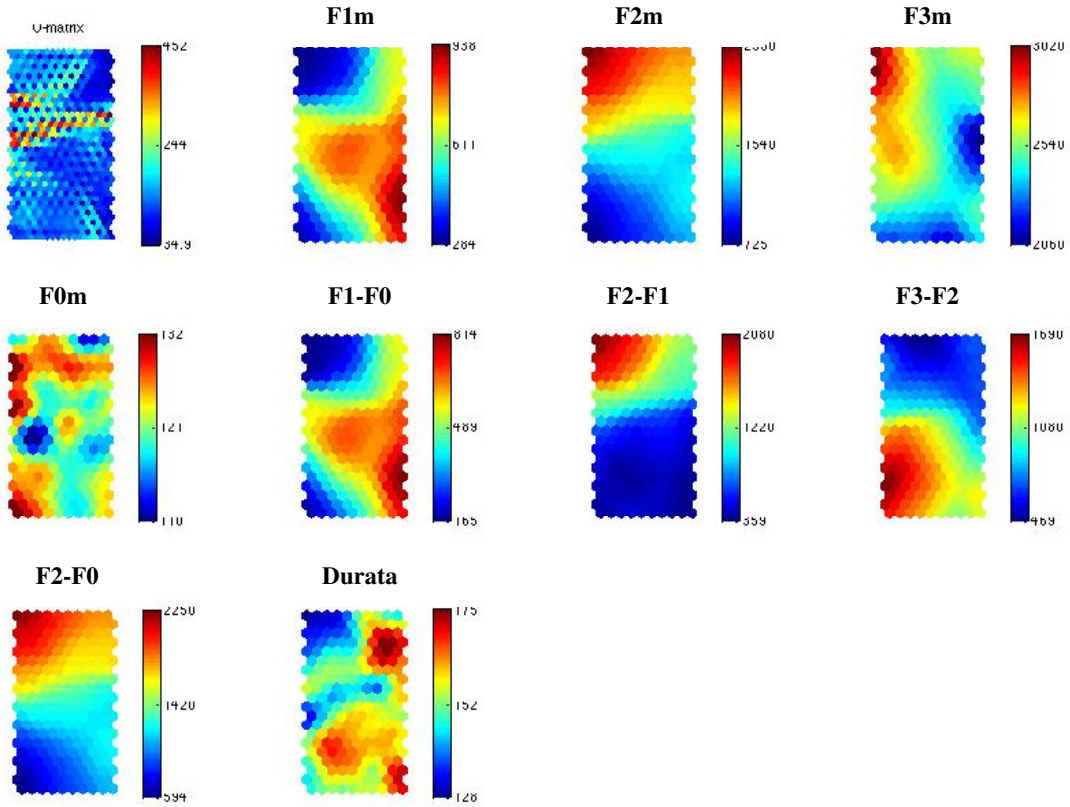

Figure 6. Visualizzazione della U-Matrix e dei valori di ciascuna delle 9 componenti delle vocali nel *corpus* TONICO PARLATO LETTO. Valori con tonalità di rosso rappresentano gradi di maggiore intensità, al contrario zone di tonalità blu definisco valori di bassa intensità

L'U-matrix calcola la distanza di ogni nodo della mappa dai propri vicini e la rappresenta graficamente come una terza dimensione, in genere costituita dalle tonalità di colore. Utilizzando tale approccio, dati vicini corrispondono a valori bassi di intensità (tonalità di blu) e i *clusters,* identificati dalla rete, sono separati dalle zone a intensità maggiore (tonalità di rosso).

La rappresentazione U-matrix della SOM per il TONICO PARLATO LETTO (già in Figura 3-a) è riportata in Figura 5. La SOM mostra due macro-classi ben distinte: in alto della mappa sono presenti le vocali /i e ɛ/ ben separate dalle restanti /a ɔ/ (occupanti la parte centrale della mappa) e da /o u/, disposte in basso a sinistra.

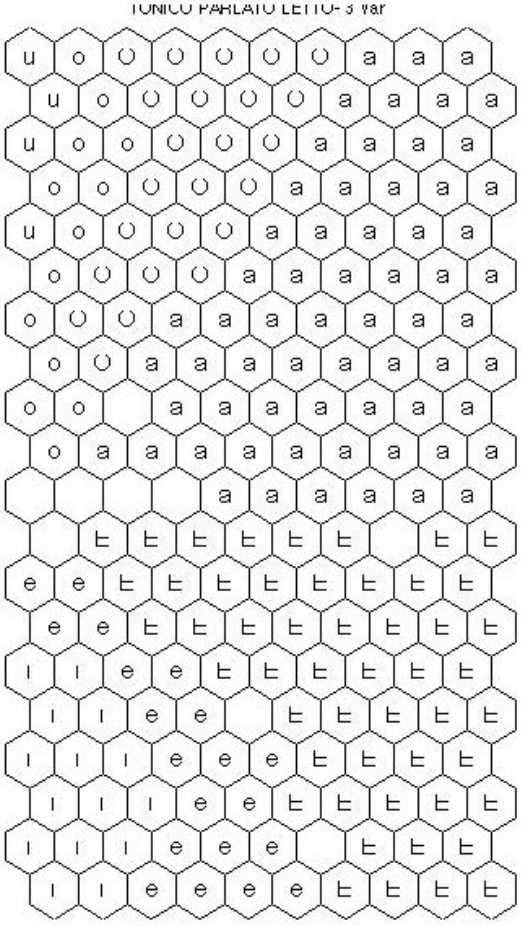

Figura 7. Organizzazione del sistema TONICO PARLATO LETTO utilizzando tre (F_1, F_2 e F_2-F_1) variabili per la rappresentazione del *set* vocalico.

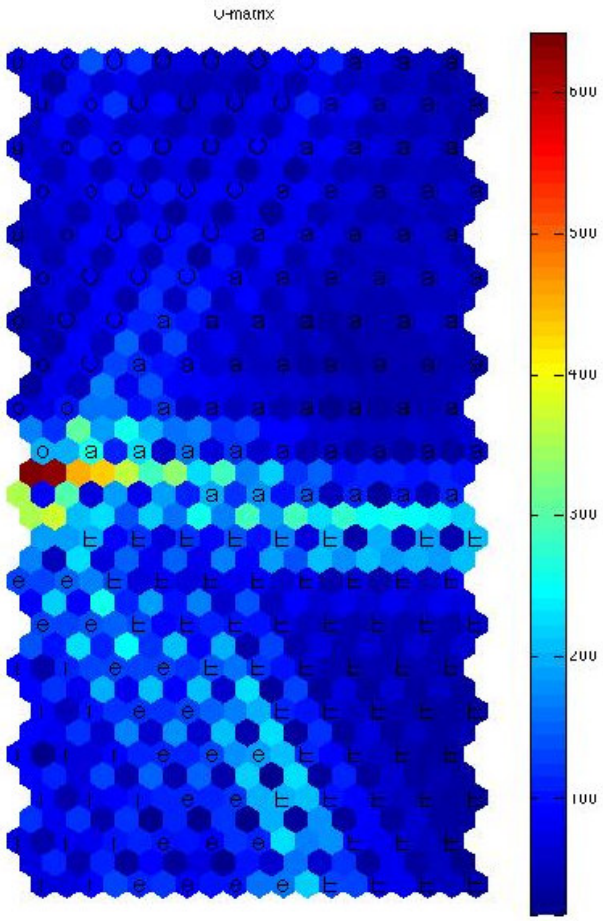

Figura 8. Visualizzazione U-matrix della SOM addestrata tramite il *corpus* TONICO PARLATO LETTO utilizzando tre (F_1, F_2 e F_2-F_1) varibili per la rappresentazione del *set* vocalico. Dati vicini corrispondono a valori bassi di intensità (tonalità di blu) e i *clusters* identificati dalla mappa sono separati dalle zone ad intensità maggiore (tonalità di rosso).

In Figura 6 sono riportati il valore di ciascuna variabile all'interno della mappa. Anche in questo caso i valori sono rappresentati tramite un *continuum* di tonalità che va dal rosso (valori elevati) al blu (valori bassi). Isolare le variabili in gioco per l'organizzazione del dato permette di confrontare visivamente più variabili tra loro alla ricerca di possibili correlazioni. Si noti, infatti, come l'area più prominente per la variabile F_2 sia quella in alto a sinistra, la stessa presente nelle variabili F_2-F_1 e F_2-F_0: in tal senso le tre variabili appaiono strettamente correlate tra loro.

Tramite comparazione visiva, inoltre, è facile identificare le variabili che maggiormente definiscono l'organizzazione finale della SOM. Si noti, infatti, come l'organizzazione delle variabili F_1, F_2 e F_2-F_1 sembrerebbe quella che maggiormente 'struttura' l'U-matrix. È interessante notare a riguardo che le zone particolarmente 'calde' di queste tre variabili coincidono, quasi in un ricalco spaziale, con alcune delle categorie spaziali della SOM finale. Ad esempio, la zona dei valori più intensi di F_1 coincide con la categoria spaziale della vocale /a/ nella SOM, mentre la zona dei valori più intensi di F_2-F_1 è sovrapponibile alla zona dedicata alle vocali /i/ ed /e/ nella SOM. In particolare F_1 ed F_2-F_1 mostrano zone di complementarietà al loro interno (si compari la zona in alto a sinistra tra le due variabili: massimamente intensa per F_2-F_1 e, al contrario, scarsamente prominente in F_1).

Al fine di verificare ulteriormente il peso di queste tre variabili all'interno di un processo di auto-organizzazione in *clusters* vocalici, una nuova serie di simulazioni è stata condotta. La nuova rappresentazione del dato vocalico prevede, in questa nuova analisi, solo tre parametri: F_1, F_2 e F_2-F_1. Il numero di neuroni e il protocollo di addestramento rimangono invariati e coincidono con quelli delle precedenti simulazioni.

Se la previsione di una possibile economia descrittiva del dato vocalico sulla base dei soli tre parametri F_1, F_2 e F_2-F_1 risultasse corretta, la SOM, ad apprendimento concluso, presenterebbe sette categorie spaziali definenti le classi naturali entro i quali le vocali sono distinte.

Le Figure 7 e 8 riportano rispettivamente la SOM finale e l'U-matrix, risultato nel nuovo processo di apprendimento. Il sistema, sulla base di tre parametri, è riuscito ad operare una coerente differenziane del *set* vocalico riuscendo a isolare e attribuire pertinenza categoriale al nostro *corpus* vocalico. Anche in questo caso valgono le osservazioni fatte in precedenza: in primo luogo, la SOM riesce a esibire una corretta categorizzazione delle vocali sulla base di un *continuum* fono-articolatorio, e la disposizione spaziale sembra seguire, in una lettura della mappa dal basso in alto, il seguente criterio: vocali anteriori /i e □/, vocale centrale /a/, vocali posteriori /□ o u/; in secondo luogo la SOM mostra un evidente isolamento di /□/, proprio del sistema tonico.

6. CONCLUSIONI

Il percorso di ricerca che abbiamo cercato di delineare nel presente lavoro può essere sinteticamente descritto come il tentativo di caratterizzare un sistema vocalico (quello pisano) all'interno di un processo di organizzazione, non-supervisionata e scevra da conoscenza aprioristica, di dati fonetici definiti sulla base di un *set* di variabili. Il processo di (auto)organizzazione è stato condotto per quattro differenti *corpora* vocalici differenziati per tipologia accentuale (vocali toniche *vs.* vocali atone) e per stile (parlato letto *vs.* parlato semispontaneo).

Come si è visto, il sistema raggiunge, ad apprendimento concluso, una configurazione finale che presenta al proprio interno coerenti categorie vocaliche organizzate spazialmente sulla base di un criterio di similarità fono-articolatoria: vocali articolarmente simili (anteriori, centrali e posteriori) sono disposte vicino all'interno delle categorie spaziali e viceversa. La configurazione finale per i quattro sistemi ha mostrato delle differenze sostanziali riguardo la bontà delle organizzazioni: una maggiore coerenza categoriale, più definita e isolata, è raggiunta per il sistema tonico e, più specificatamente, per lo stile più controllato (il parlato letto).

È stato compiuto anche un tentativo di 'economia descrittiva' del sistema vocalico. L'utilizzo di una metodologia largamente usata nella letteratura delle SOM, l'U-matrix, ha

permesso inoltre di evidenziare visivamente, e nel dettaglio, le categorie vocaliche emerse dal processo di organizzazione e al contempo ha consentito lo studio delle variabili che maggiormente hanno pesato per l'individuazione finale delle categorie vocaliche. Abbiamo registrato il valore di ciascuna delle nove variabili all'interno del processo di organizzazione e proceduto a una nuova serie di addestramento riducendo il vettore rappresentante ogni vocale da nove a tre componenti. Il sistema ha mostrato, anche in questa occasione, un'organizzazione del *corpus* in categorie vocaliche rispettando altresì il criterio di similarità fono-articolatoria riscontrato precedentemente. La scelta della rete cade naturalmente sulle componenti acustiche che meglio individuano e categorizzano i sistemi vocalici: la prima formante, la seconda formante, la differenze spettrale F_2-F_1.

7. BIBLIOGRAFIA

Bertinetto, P.M. (2001) (editor), *AVIP – Archivio di Varietà di Italiano Parlato*, 4 CD-Rom, Ufficio Pubblicazioni della Classe di Lettere della Scuola Normale Superiore, Pisa.

Calamai, S. (2004), *Il vocalismo tonico pisano e livornese. Aspetti storici, percettivi, acustici*, Edizioni dell'Orso, Alessandria.

Calamai, S., Agonigi, M., Ricci I. (2005), Metodologie statistiche di classificazione a confronto: analisi discriminante e CART, in *Misura dei parametri: aspetti tecnologici ed implicazioni nei modelli linguistici* (P. Cosi, editor), I Convegno Nazionale Associazione Italiana di Scienze della Voce, Padova, 2-4.XII.2004, Padova, EDK: 619-649.

Carlson, R., Granstrom, B. & Fant, G. (1970), Some studies concerning perception of isolated vowels, *STL-QPSR*, 2-3, 19-35.

Cioni, L. (2001), Multi-Speech e ESPS: tecniche di scripting per l'analisi acustica, *Quaderni del Laboratorio di Linguistica della Scuola Normale Superiore*, 2, 125-137.

Cosi, P. (1991), Riconoscimento di sequenze vocaliche tramite reti neurali, in *L'interfaccia tra fonologia e fonetica.* Atti del Convegno di Padova, 15.XII.1989 (E. Magno Caldognetto & P. Benincà, editors), Unipress, Padova, 165-172.

Delattre, P. C., Liberman, A. M., Cooper, F. S., & Gerstman, L. J. (1952), An experimental study of the acoustic determinants of vowel color; observations on one- and two-formant vowels synthesized from spectrographic patterns, *Word*, 8, 195-210.

Di Benedetto, M.-G. (1991), Complex relation between F1 and *F*0 in determining vowel height: acoustic and perceptual evidence, *Studi Italiani di Linguistica Teorica ed Applicata* 20, 579-603.

Di Benedetto, M.-G. (1994), Acoustic and perceptual evidence of a complex relation between F1 and *F*0 in determining vowel height, *Journal of Phonetics* 22, 205-224.

Hirahara, T. & Kato, H. (1992), The effect of *F*0 on vowel identification, in *Speech Perception, Production and Linguistic Structure* (Y. Tohkura, E. Vatikiotis-Bateson & Y. Sagisaka, editors), Tokyo & Amsterdam, Ohmsha & IOS Press, 89-112.

Jassem, W. & Grygiel, W. (2004), Off-line classification of Polish vowel spectra using artificial neural networks*, Journal of the International Phonetic Association*, 34, 37-52.

Johnson, K. (1990), Contrast and normalization in vowel perception, *Journal of Phonetics*, 18, 229-254.

Kohonen, T. (2001), *Self-Organizing Map*, Springer, Berlin.

Traunmüller, H. (1981), Perceptual dimension of openness in vowels, *JASA*, 69, 1465-1475.

Ultsch, A. & Siemon, H. P. (1990), Kohonen's self organizing feature maps for exploratory data analysis, in *Proceedings of International Neural Network Conference*, Dordrecht, The Netherlands, 305-308.