RAPPRESENTAZIONI ACUSTICHE E UDITIVE DELLE VOCALI ITALIANE

P. Cosi, F. Ferrero e K. Vagges

Centro di Studio per le Ricerche di Fonetica (CNR)
Via Anghinoni, 10 - 35121 Padova (ITALY)
Tel.: (+39) 49 8274418
FAX: (+39) 49 8274416
EMail: COSI@CSRF00.CSRF.PD.CNR.IT

SOMMARIO

Per caratterizzare le vocali italiane viene fatto comunemente riferimento ai loro diagrammi di esistenza definiti nello spazio F1-F2. In questo lavoro vengono analizzati nuovi spazi rappresentativi e viene considerato un nuovo e ben più ampio materiale di analisi. In particolare, è stata analizzata statisticamente la capacità discriminativa di alcune rappresentazioni acustiche e uditive delle vocali presenti in un sottoinsieme, costituito interamente da bisillabi senza senso del tipo CV'CV pronunciati da 20 soggetti maschili e 20 soggetti femminili, del nuovo *data-base* denominato AIDA.

INTRODUZIONE

Volendo caratterizzare i fonemi vocalici dell'italiano ci si riferisce essenzialmente ai loro diagrammi di esistenza definiti nello spazio F1-F2 [1]. Non essendo ancora apparso in letteratura un aggiornamento completo ed affidabile dei risultati ottenuti in quello studio ci è sembrato interessante rianalizzare le vocali, considerando, come materiale di analisi, un più ampio e completo data-base, rispetto a quello utilizzato nello studio sopra citato e utilizzando inoltre rappresentazioni acustiche e uditive già considerate in precedenti studi [2]. Recentemente è stato sviluppato dalla Commissione Nazionale Basi Dati Vocali Italiana, promossa ad-hoc dall'Istituto Superiore delle Poste e Telecomunicazioni, un nuovo data-base vocale denominato AIDA [3] per il riconoscimento automatico della parola sia *speaker-dipendente* (SD) che *speaker-*

imdipendente (SI). Essendo AIDA, un insieme di dati vocali distribuibile a tutti coloro che abbiano necessità e convenienza ad utilizzare dei dati comuni, esigenza sempre più sentita a livello internazionale, sia a livello di ricerca, che in compiti più formali, quali ad esempio le omologazioni, ci è sembrato importante, vista anche la sua completezza, utilizzarlo come materiale di base per questo studio sulle vocali.

Il CORPUS AIDA

Il corpus AIDA, è diviso in due data-base: uno speaker-dipendente ed uno speaker-indipendente. Ognuno dei due data-base è contenuto in tre CD-ROM. Per le registrazioni sono stati selezionati 40 parlatori (20 maschi e 20 femmine), scelti, in parte a Torino ed in parte a Roma, sulla base di un criterio di eguale distribuzione sul territorio nazionale. Tutti i parlatori hanno pronunciato una volta il materiale vocale per la parte SI, mentre otto soggetti, selezionati tra i 40, hanno poi ripetuto altre 5 volte la registrazione dello stesso materiale per la parte SD. Per quanto riguarda il materiale registrato, sono presenti: un brano di calibrazione, bisillabi CV'/ta/ e /t/V'CV (essendo C e V tutte le possibili consonanti e vocali dell'italiano), i principali cluster bi- e tri-consonantici iniziali ed intervocalici e le cifre da 0 a 9. In questo lavoro sono state considerate esclusivamente le vocali estratte dai bisillabi con variazione di consonante iniziale ed intervocalica. I segnali contenuti in AIDA sono stati campionati, dopo un opportuno filtraggio *anti-aliasing*, a 20 kHz.

METODOLOGIA DI ANALISI

Vista l'enormità del materiale vocale da analizzare (18.880 foni vocalici), si sono dovute progettare opportune tecniche automatizzate di analisi per l'individuazione delle zone vocaliche e per l'elaborazione dei parametri di interesse. In particolare, mediante gli algoritmi AMDF [4] e LPC [5], nella regione di massima energia di ogni vocale bersaglio, sono stati misurati, ogni 5ms e utilizzando finestre di analisi di 300ms, i seguenti parametri:

f0	-	frequenza fondamentale
Fn (n=1,2,3)	-	frequenze formantiche
Bn (n=1,2,3)	-	larghezze di banda
Ln (n=1,2,3)	-	intensità (dB).

Le formanti e le larghezze di banda sono state estratte calcolando le radici del denominatore del filtro LPC. Successivamente, allo scopo di validarne l'affidabilità, le formanti così ottenute sono state paragonate ai valori calcolabili direttamente misurando i punti di massimo del ritardo di gruppo dello spettro LPC, dimostrando una notevole concordanza. Per quanto riguarda l'individuazione delle regioni di massima energia, la tecnica automatica utilizzata è essenzialmente basata sull'energia a breve termine del segnale vocale. Come illustrato in Figura 1, mediante una soglia variabile decrescente, vengono localizzate le due regioni ad energia più elevata. Successivamente viene

calcolata la media dei parametri di interesse all'interno di una finestra di 7 frame centrata sul frame corrispondente alla massima energia per entrambe le regioni (m1, m2).

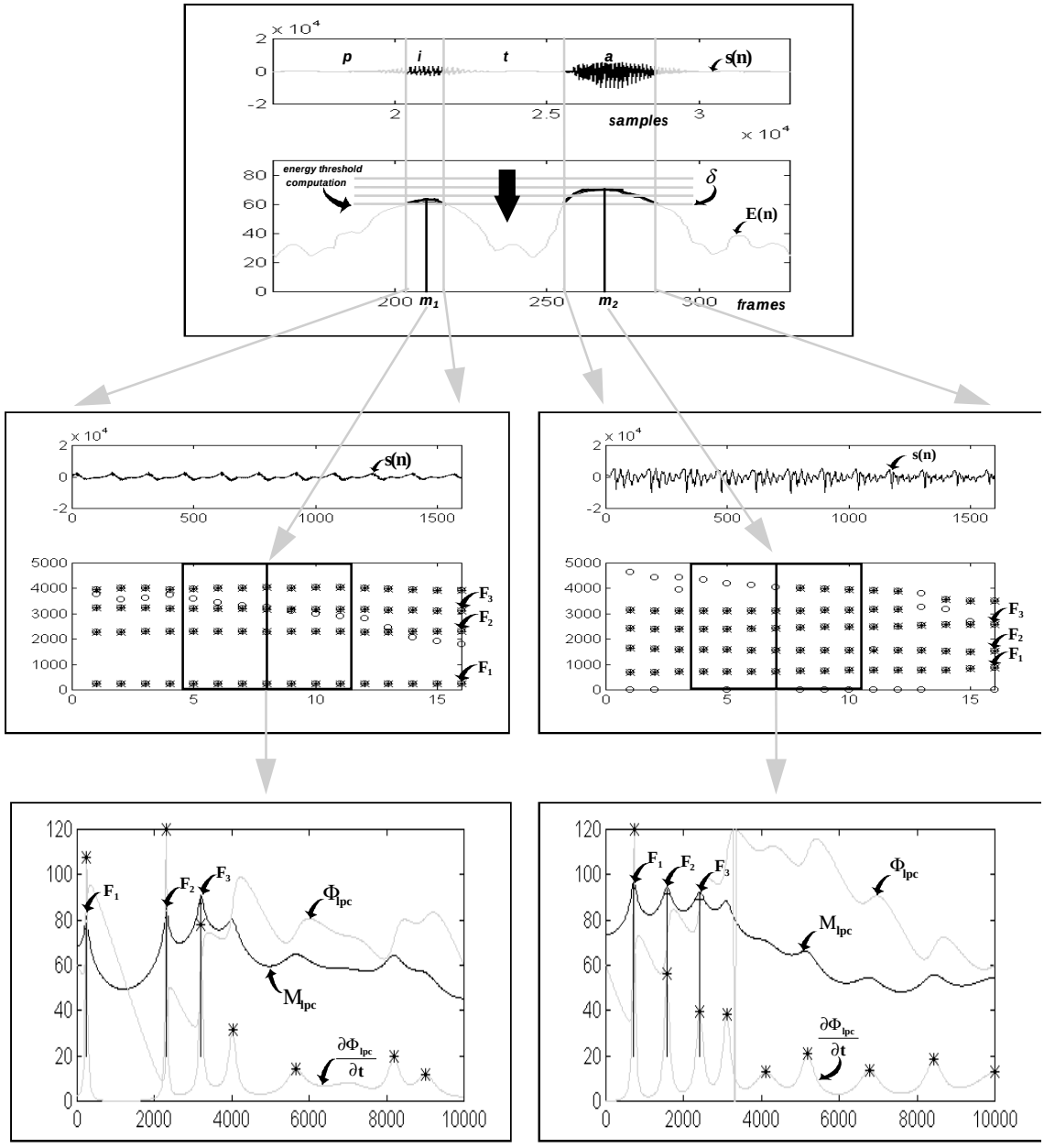

Figura 1. Localizzazione automatica dei foni vocalici mediante variazione della soglia di energia (δ) ed estrazione automatica delle formanti tramite il metodo LPC, mediando i valori corrispondenti ad una finestra di 7 frame centrata sul frame relativo al massimo di energia, per entrambe le vocali target. In figura sono indicati modulo (M) e fase (Φ) dello spettro LPC ed il corrispondente ritardo di gruppo ($\partial\Phi$ / ∂t) utilizzato solo a scopi di controllo di affidabilità delle misure formantiche.

Allo scopo di escludere eventuali errori di classificazione, sempre possibili in una procedura automatica, prima di procedere alle elaborazioni statistiche è stata effettuata una selezione degli stimoli sulla base di un criterio di appartenenza a regioni ammissibili per ogni valore formantico.

RAPPRESENTAZIONI ACUSTICHE E UDITIVE

Oltre allo spazio classico (F1,F2,F3), sono stati considerati altri spazi rappresentativi allo scopo di determinarne statisticamente l'eventuale diversa capacità discriminativa.

Complessivamente sono state considerate le seguenti rappresentazioni:

Σ_1 :	F1, F2, F3	(Hz)
Σ_2 :	F1-f0, F2-F1, F3-F2	(Hz)
Σ_3 :	F1, F2, F3	(Bark)
Σ_4 :	F1-f0, F2-F1, F3-F2	(Bark)

Per quanto riguarda la trasformazione Hz/Bark la formula utilizzata è stata quella proposta da Schroeder et al. (1979):

$$\mathbf{b = 7 \ln\left\{\left(\frac{f}{650}\right) + \left[\left(\frac{f}{650}\right)^2 + 1\right]^{\frac{1}{2}}\right\}} \qquad \mathbf{(1)}$$

In Figura 2 sono illustrati i diagrammi di esistenza delle vocali maschili e femminili, nelle varie rappresentazioni, relativamente alla prima vocale degli stimoli CV'/ta/. Analoghi diagrammi sono stati costruiti per le vocali finali degli stimoli /t/V'CV. Le ellissi di dispersione sono calcolate con un criterio di copertura dello spazio totale delle variabili in esame, in particolare la rappresentazione di F1 e F2 negli spazi Σ_{1-4}, del 75%. In altre parole il 75% degli stimoli presenta valori formantici contenuti nelle varie ellissi. In Figura 3 sono illustrati i diagrammi di esistenza, questa volta considerati nelle coordinate [F3-F2, F1-f0] e [F3-F2, F2-F1] dello spazio Σ_4, delle stesse vocali utilizzate in Figura 2. Da questa figura risulta evidente l'elevato grado di potere discriminante fra vocali anteriori e posteriori ottenibile in tale rappresentazione. In Tabella 1 sono riassunti i valori delle medie e delle deviazioni standard ottenuti per le prime tre formanti nello spazio Σ_1, sia per le vocali maschili che per quelle femminili.

ANALISI STATISTICA

Allo scopo di valutare la rappresentazione migliore per una più affidabile discriminazione vocalica è stata applicata l'analisi discriminante lineare (*linear discriminant analysis).* Con tale analisi, infatti, viene determinata una combinazione lineare dei parametri rappresentativi degli stimoli in esame tale da rendere massima la distinzione tra i valori medi dei gruppi, nel nostro caso tra le vocali. Analizzando le risultanti matrici di confusione relative agli spazi Σ_{1-4}, rappresentate nella Tabella 2, si può concludere che non vi è una significativa differenza fra gli spazi considerati, sottolineando, tuttavia, che le migliori performance discriminative sono ottenibili

utilizzando lo spazio Σ_3. Si nota, inoltre, un lieve peggioramento delle prestazioni utilizzando le due rappresentazioni normalizzate rispetto a f0 (Σ_3, Σ_4). Rispetto ai risultati ottenuti in [1], vi è una significativa differenza riguardo al dominio di esistenza delle vocali posteriori femminili, evidenziata soprattutto dalla vocale /ɔ/, il cui *range* di frequenze risulta sensibilmente ridotto nella dimensione F2.

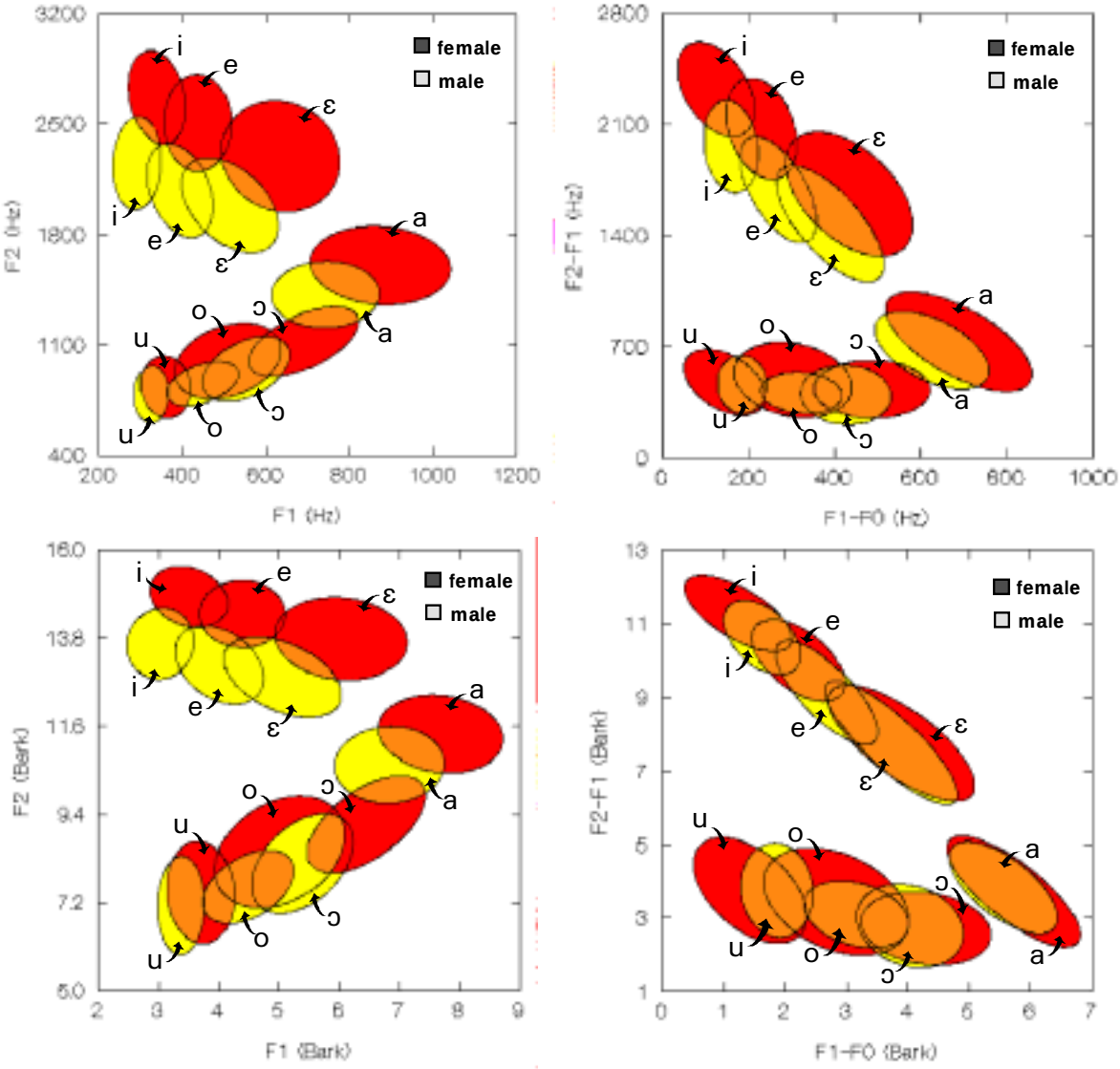

Figura 2. Diagrammi di esistenza delle vocali maschili e femminili relativi alla prima vocale degli stimoli CV/ta/ del data-base AIDA. Sono utilizzate le quattro rappresentazioni Σ_{1-4} indicate nel testo.

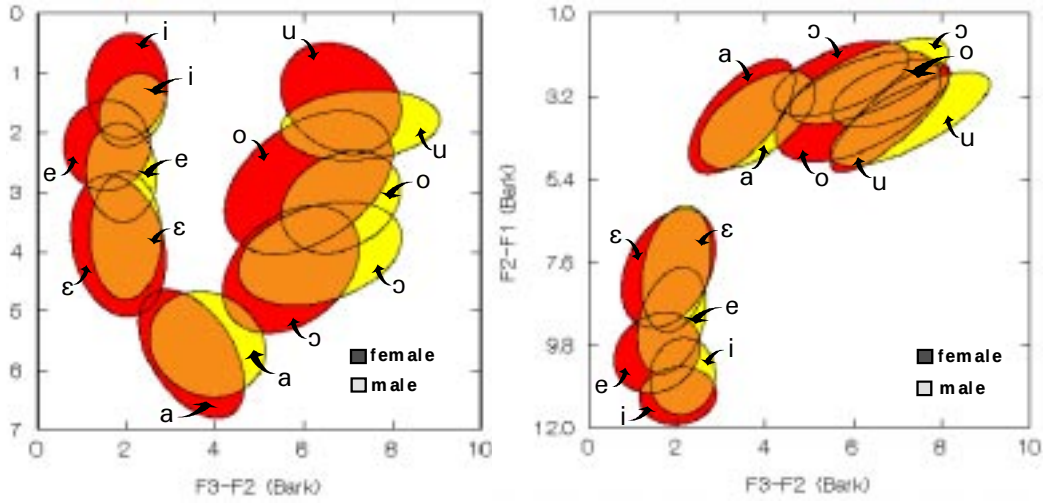

Figura 3. Diagrammi di esistenza nelle coordinate [F3-F2, F1-f0] e [F3-F2, F2-F1] dello spazio Σ_4 delle vocali maschili e femminili relativi alla prima vocale degli stimoli CV/ta/ del data-base AIDA.

Σ_1

	a	e	ɛ	i	o	ɔ	u	
a	95.4	0.0	1.4	0.0	0.0	3.2	0.0	
e	0.0	68.8	13.9	17.0	0.0	0.0	0.3	
ɛ	1.0	22.3	74.8	1.8	0.0	0.0	0.0	
i	0.0	12.7	0.0	87.3	0.0	0.0	0.0	
o	0.0	0.0	0.0	0.0	74.3	15.8	9.9	
ɔ	3.4	0.7	2.1	0.7	21.6	71.3	0.2	
u	0.0	0.0	0.0	0.0	0.6	0.0	99.4	
							m1	81.6
							m2	92.1

Σ_2

	a	e	ɛ	i	o	ɔ	u	
a	95.4	0.0	1.7	0.0	0.0	2.9	0.0	
e	0.0	63.9	17.3	18.6	0.0	0.0	0.3	
ɛ	0.3	22.1	74.6	3.1	0.0	0.0	0.0	
i	0.0	13.0	0.0	87.0	0.0	0.0	0.0	
o	0.0	0.0	0.0	0.0	73.0	17.8	9.3	
ɔ	2.5	0.7	2.1	0.7	21.0	72.9	0.2	
u	0.0	0.0	0.0	0.0	2.4	0.0	97.6	
							m1	80.6
							m2	91.8

Σ_3

	a	e	ɛ	i	o	ɔ	u	
a	96.8	0.0	1.4	0.0	0.0	1.7	0.0	
e	0.0	74.5	11.6	13.7	0.3	0.0	0.0	
ɛ	0.8	21.8	76.9	0.5	0.0	0.0	0.0	
i	0.0	10.1	0.0	89.9	0.0	0.0	0.0	
o	0.0	0.0	0.0	0.0	73.6	17.1	9.3	
ɔ	4.1	0.9	2.5	0.5	20.1	72.0	0.0	
u	0.0	0.0	0.0	0.0	1.2	0.0	98.8	
							m1	83.2
							m2	93.3

Σ_4

	a	e	ɛ	i	o	ɔ	u	
a	96.3	0.0	1.4	0.0	0.0	2.3	0.0	
e	0.0	70.4	15.0	14.4	0.0	0.0	0.3	
ɛ	0.5	22.9	75.8	0.8	0.0	0.0	0.0	
i	0.0	10.6	0.0	89.4	0.0	0.0	0.0	
o	0.0	0.0	0.0	0.0	70.9	19.5	9.6	
ɔ	3.6	0.9	2.5	0.7	20.3	71.8	0.2	
u	0.0	0.0	0.0	0.0	3.1	0.0	97.0	
							m1	81.6
							m2	92.7

Tabella 1. Valori delle medie ottenuti per le prime tre formanti nello spazio Σ_1, sia per le vocali maschili che per quelle femminili.

Male	i	e	ɛ	a	o	ɔ	u
F1	291	394	513	742	552	447	325
F2	2251	2082	1989	1420	949	856	789
F3	3079	2752	2669	2532	2569	2528	2529

Female	i	e	ɛ	a	o	ɔ	u
F1	339	436	630	875	688	506	360
F2	2672	2508	2302	1614	1115	990	838
F3	3595	3158	2999	2697	2712	2606	2466

Tabella 2. Matrici di confusione (%), relative all'analisi discriminante lineare negli spazi Σ_{1-4}. m1 ed m2 indicano rispettivamente la media delle classificazioni corrette utilizzando tutte e sette le vocali oppure soltanto 5 vocali, raggruppando assieme le coppie aperte e chiuse delle vocali /e/ ed /o/.

BIBLIOGRAFIA

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

[2] Di Benedetto M. G., Flammia G., *Vowel Distinction Along Auditory Dimension: a Comparison between a Statistical and a Neural Classifier*, Proc. of VERBA-90, Rome, 1990, pp. 248-255.

[3] Castagneri G., Vagges K., *The Italian National Data-Base for Speech Recognition*, Proc. ICSLP-90, Kobe (Japan), Nov. 18/22, 1990, pp. 1285-1287.

[4] Ross M.J., Shaffer H.L., Cohen A., Freudberg R. and Manley H.J., *Average Magnitude Difference Function Pitch Extractor*, IEEE Transaction on Aoustics Speech and Signal Processing. Vol. ASSP-22, Oct. 1974, pp.353-362.

[5] Markel J.D., Gray A.H., *Linear Prediction of Speech*, Springer-Verlag, Berlin 1976.