SPERIMENTAZIONE DI SISTEMI SENZA ADDESTRAMENTO PER L'ESTRAZIONE DI FONEMI VOCALICI

Francesco Borchi, Monica Carfagni, Matteo Nunziati
Università degli Studi di Firenze, Dipartimento di Meccanica e Tecnologie Industriali (DMTI)
{francesco.borchi,monica.carfagni,matteo.nunziati}@unifi.it

1. SOMMARIO

L'analisi di corpora vocali è una delle attività più importanti sia per il campo della ricerca che per quello delle attività commerciali legate all'analisi della voce. In entrambi i casi, la possibilità di analizzare ed impiegare database sufficientemente ampi (tali da essere considerate rappresentative dei fenomeni di interesse) è un tassello fondamentale per la correttezza dei risultati prodotti. Tuttavia, il trattamento dei dati contenuti nei corpora vocali, compiuto al fine di estrarre tutte e sole le informazioni di interesse per una data attività, richiede molte ore di lavoro.

Nell'arco degli anni sono stati sviluppati sistemi automatici di segmentazione, che fanno uso di modelli matematici complessi (vedi i sistemi HMM in Huang *et alii* (2001)): tali modelli richiedono, a loro volta, una mole considerevole di dati per la loro messa a punto ed addestramento e si presenta il problema di reperire strumenti intermedi, che aiutino nello sviluppo.

Nell'ambito di un progetto di ricerca relativo al riconoscimento automatico del parlante, il Dipartimento di Meccanica e Tecnologie Industriali dell'Università di Firenze (di seguito DMTI) ha iniziato lo sviluppo e la sperimentazione di un sistema specializzato per la catalogazione di fonemi vocalici accentati. Uno degli obiettivi per lo sviluppo di tale sistema è stato quello di realizzare uno strumento totalmente automatico, che richiedesse costi di taratura minimi, possibilmente nulli. Un altro obiettivo dello studio è da ricercarsi nella possibilità di trasferimento tecnologico dello strumento, che dovrà risultare impiegabile da chiunque anche a scopo commerciale, senza la necessità di adottare un sistema di segmentazione dai costi elevati. A tal fine, il lavoro si è orientato sullo sviluppo di due strumenti diversi: il primo, derivante dalla generalizzazione del metodo proposto da Cosi *et alii* (1995), impiegato per definire la posizione delle vocali all'interno dell'enunciato; il secondo, caratterizzato dall'utilizzo del piano formantico F1-F2 (in Hz), impiegato per definire il triangolo vocalico di un parlante generico e per classificare le diverse vocali accentate individuabili nel parlato.

Allo stato attuale, un prototipo funzionante di tale sistema è sotto analisi presso il dipartimento. Il prototipo, nella configurazione attuale, è stato testato solo in condizioni testo-dipendenti su parlato letto in laboratorio.

2. IL METODO IBT-DESF

L'idea di fondo del sistema, che verrà descritto di seguito, è scaturita dalla lettura del lavoro condotto da Cosi *et alii* (1995), in cui è stato sviluppato un sistema automatico di individuazione di fonemi vocalici. Le vocali oggetto di studio erano quelle contenute nel corpus AIDA. Gli enunciati presi in considerazione in tale lavoro erano costituiti da bisillabi di tipo CV'CV, per tale motivo erano note agli autori sia le posizioni delle vocali accentate all'interno delle singole sequenze di dati, sia la loro classificazione. Il metodo presentato al convegno AIA del 1995 era basato sull'analisi dell'andamento dell'energia a

breve termine del segnale. L'idea di fondo era la seguente: dal momento che le vocali sono realizzate ostacolando in maniera minima il flusso d'aria, mentre caratteristica delle consonanti è la forte occlusione dell'apparato fonatorio, il livello di energia associato al segnale deve subire un incremento in prossimità delle suddette vocali.

A partire da questa considerazione, di volta in volta, veniva definita una soglia decrescente, il cui scopo era quello di isolare le due aree di massima energia associate alle vocali dei bisillabi. Individuate tali aree, era immediata la loro classificazione: sia l'accento sia il tipo di vocale erano deducibili dalla trascrizione associata al file audio. Una volta raccolti i dati sui fonemi, veniva valutata la capacità di discriminazione fornita da diverse rappresentazioni su piani formantici. A partire dalle intuizioni di tale lavoro, nel nostro studio abbiamo tentato di estendere il campo di impiego della metodologia al fine di individuare le aree di massima intensità associate alle vocali presenti in un testo generico.

Il metodo indicato da Cosi *et alii* (1995) si basava su un'ipotesi ben precisa: ogni bisillabo doveva contenere due picchi di energia, ognuno dei quali associato ad una vocale. Tale ipotesi decade nel contesto della presente ricerca, dal momento che il parlato non viene considerato noto a priori e dal momento che eventuali picchi di energia possono essere dovuti anche ad eventi eccezionali. Tale incremento rimane comunque una condizione necessaria alla presenza di fonemi vocalici: a partire da questa considerazione è stato sviluppato un selettore qui definito IBT-DESF (Intensità a Breve Termine – Durata, Esistenza e Stabilità delle Frequenze caratteristiche). Nei paragrafi che seguono verranno descritte le caratteristiche di tale soluzione.

2.1 Il selettore IBT

Lo scopo del selettore IBT è quello di individuare il punti di massimo dell'intensità a breve termine (IBT) ed isolarle dal contesto. In questo caso, è stata definita una soglia $\boldsymbol{I_{min}}$, tramite la quale discriminare ciò che è considerabile come un fonema vocalico accentato e ciò che non lo è (vedi figura 1). Ovviamente in casi estremi come in chiusura di frase, ove spesso l'intensità ha un decremento generalizzato, le vocali sono inesorabilmente scartate, in quanto non riescono a superare la soglia minima di intensità imposta dal sistema. Tuttavia, questo comportamento non è stato considerato scorretto: fenomeni in chiusura di frase, associati a cali diffusi di intensità, sono spesso associati anche ad una fonazione non standard da parte del parlatore. Dal momento che lo scopo del metodo è quello di raccogliere informazioni sul contenuto formantico dei fonemi vocalici, si è ritenuto corretto escludere dalle statistiche i casi estremi.

Risulta evidente che la risposta del metodo IBT dipende dalla soglia di intensità scelta dall'operatore. A sua volta tale valore deve essere associato al livello del segnale acquisito. Al fine di verificare il sistema nel suo complesso, si è preferito, in questa fase, limitare le variabili in gioco. Per tale motivo, si sono realizzate delle registrazioni pilota di livello pressoché omogeneo (descritte nel dettaglio in seguito) e si è imposta una soglia fissa di 60 dB, che, in questa sede, non entra nel problema come variabile ma come costante nota. In termini generali si può comunque dire che: dato l'insieme $\boldsymbol{M}$ dei possibili massimi locali di intensità, il metodo IBT è in grado di isolare tra questi un sottoinsieme $\boldsymbol{N}$ di picchi aventi intensità maggiore di quella imposta dall'operatore.

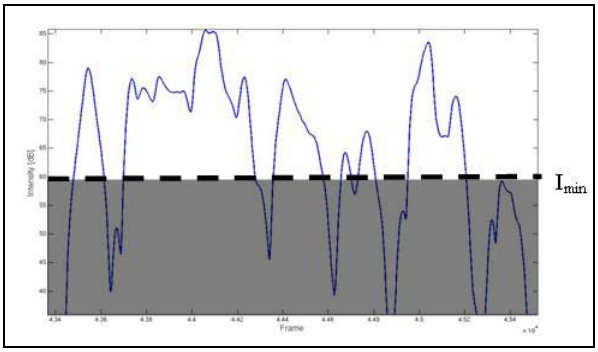

Figura 1: schema esemplificativo del comportamento selettivo della soglia $\boldsymbol{I_{min}}$. I massimi locali giacenti nella zona annerita sono scartati dal selettore.

2.2 Il selettore DESF

Dal momento che il valore di IBT non è da solo sufficiente a rivelare la presenza di una vocale, il metodo è stato esteso includendo altri parametri di selezione. Prima fra tutti la durata (D) del picco di intensità. Dal momento che possibili eventi eccezionali possono inserirsi nell'insieme ***N***, la valutazione della durata del picco permette una prima discriminazione. Anche in questo caso si ricorre a condizioni necessarie ma non sufficienti: di fatto, viene imposta una durata minima $\boldsymbol{D_{min}}$, che, se non rispettata, porta allo scarto dei massimi locali. Risulta evidente come una soglia di questo tipo sia necessariamente legata al ritmo di eloquio di ogni soggetto. Tuttavia la sperimentazione ha dimostrato che esiste una fascia di pochi millisecondi (30 ms.), indipendente dal parlante e sufficientemente ampia da garantire un'eliminazione dei fenomeni spuri. Le prove effettuate hanno dimostrato che al di sotto di tale soglia si presentano fenomeni di massimo locale relativi ad eventi spuri o, al più, massimi locali relativi ad eventi fonatori non sufficientemente intensi da poter essere associati a vocali accentate. In questo caso, infatti, la selezione IBT porta a conservare solo una parte ridotta (quella più ricca di energia) della fase di tenuta di possibili vocali, riducendo di molto la durata del picco in esame (al fine di comprendere meglio il comportamento combinato delle due soglie si faccia riferimento allo schema esemplificativo di figura 2). In questa fase viene tra l'altro scartata una parte rilevante dei fonemi nasali rimanenti dopo la prima selezione. Data la fase di sviluppo del sistema, non è possibile al momento quantificare con una percentuale precisa questo fenomeno, che risulta tuttavia evidente in tutte le tracce sin'ora analizzate.

Individuato così un nuovo sottoinsieme ***V***, incluso in ***N***, sono state introdotte altre grandezze di controllo, relative alle caratteristiche fonetiche dei picchi. Ogni vocale è classificabile come un fonema sonoro, perciò uno degli elementi di rilievo per la discriminazione è senza dubbio la presenza nel segnale della frequenza fondamentale (di seguito f0). Le vocali accentate sono inoltre caratterizzate da una durata tendenzialmente maggiore rispetto agli altri fonemi vocalici, come confermato da Alfano (2005): data la maggiore durata del fonema, si è ritenuto che la fase di tenuta consentisse una maggiore stabilizzazione dei valori in frequenza. Questo ha condotto ad ipotizzare che fosse possibile discriminare le vocali tramite vincoli di stabilità imposti sia alla f0 sia alle frequenze formanti[1] (d'ora in avanti f1 ed f2). In sintesi, si è tentato di discriminare i fonemi vocalici

[1] Si è limitato lo studio alle prime due formanti.

accentati tramite l'analisi dell'esistenza e della stabilità delle frequenze caratteristiche (ESF).

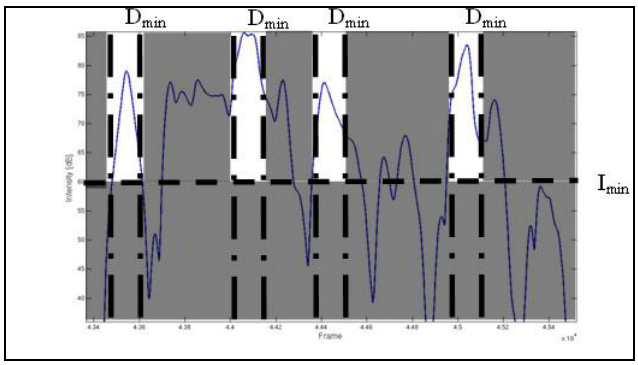

Figura 2: schema esemplificativo del comportamento selettivo delle soglie $\boldsymbol{I_{min}}$ e $\boldsymbol{D_{min}}$. La zona annerita della figura rappresenta i massimi locali di intensità scartati dai selettori.

Il limite di questo approccio consiste nella necessità di accumulare un set $\boldsymbol{Z}$ di dati, sufficiente a definire in termini numerici il concetto di stabilità. In questa sede, al fine di costituire tale set, è stata presa in considerazione una piccola quantità di vocali tratta da dati segmentati manualmente. Tale quantità è associata ad una porzione dei fonemi, presenti in quattro tracce appartenenti a parlanti diversi e ricavate dal piccolo corpus realizzato per il progetto pilota qui descritto (il corpus sarà descritto dettagliatamente più avanti).

Per il selettore non è rilevante conoscere il valore esatto delle frequenze caratteristiche, quanto la loro variazione del tempo. Per definire il livello di stabilità delle frequenze viene considerato un intervallo ΔT di 30 ms.. centrato rispetto al picco di intensità individuato dall'IBT. Per ogni singolo fonema, all'interno dell'intervallo ΔT, vengono estratte f0, f1 ed f2.

Ottenuti i dati grezzi, per ogni campione z_h appartenente a $\boldsymbol{Z}$, è stata calcolata la media μ_{ih}, dove i=0,1,2 indica una delle frequenze caratteristiche. È stata inoltre calcolata la media sull'intero set $\boldsymbol{Z}$, indicata con μ_i, e la deviazione standard, indicata con σ_i. In altre parole, i dati in frequenza f_{ihj}, dove j=1,…,30 indica l'istante di calcolo, sono stati considerati come possibili valori di una stessa variabile aleatoria $\boldsymbol{F_i}$ con distribuzione normale $N_i=N(\mu_i, \sigma_i)$. In questa sede sono stati trascurati per semplicità i possibili fenomeni di covarianza.

Successivamente, sono state valutate le differenze istantanee tra i tracciati f_{ihj} e le medie μ_{ih} di ogni singolo fonema z_h. Ognuna delle vocali è associata quindi a 90 valori istantanei di variazione (uno per ogni millisecondo e per ogni frequenza). Tali valori saranno di seguito indicati con Δf_{ihj}, mentre con $|\Delta f|_{ih,max}$ si farà riferimento al massimo del loro valore assoluto. Infine, è stata definita una soglia di stabilità convenzionale.

La soglia è valutata a partire da un coefficiente reale positivo γ. Tale coefficiente rappresenta il livello di stabilità imposto dall'operatore al fine di poter includere un fonema tra i possibili candidati delle vocali accentate. Per la stima di γ sono state considerate le seguenti relazioni fortemente semplificate:

(1a,1b, 1c)
$$\begin{cases} P_{tot,h} = \theta \\ P_{i,h} = P\left(\left|\Delta f\right|_{ih,\max} \le \gamma\sigma_i\right) \\ P_{tot,h} = \prod_{i=0}^{2} P_{i,h} \; con \; P_{i,h} = P_{i+1,h} \; \forall(i,h) \end{cases}$$

Dove θ è una variabile di progetto, che indica la percentuale di fonemi appartenenti a ***Z*** che si vuol conservare. Come conseguenza diretta delle precedenti relazioni, si deduce che è possibile ricavare γ, imponendo:

(2)
$$P_{i,h} = \sqrt[3]{\theta}$$

In questa sede, θ è stato posto uguale a 0.99, al fine di preservare la quasi totalità dei campioni del set ***Z*** (d'altro canto tali fonemi erano interamente costituiti da vocali accentate segmentate manualmente).

Fissata la soglia di selezione tramite la (2), il selettore ripete l'analisi su i generici picchi di intensità inclusi nell'insieme ***V***: vengono calcolate quindi le differenze istantanee Δf_{ilj} tra i tracciati di ogni singolo fonema v_l dell'insieme ***V*** e le relative medie μ_{il}. In accordo con l'equazione (2) e con quanto detto in precedenza relativamente alla natura dei fonemi vocalici, i fonemi dell'insieme ***V*** sono accettati se e solo se risultano verificati i seguenti vincoli:

(3)
$$\begin{cases} \exists f_{ilj} \quad \forall(i,j) \\ \left|\Delta f\right|_{il,\max} = \max_{j=1,\ldots,30}\left|f_{ilj} - \mu_{il}\right| \le \gamma\sigma_i \quad \forall(i,j) \end{cases}$$

A titolo illustrativo, si riporta in figura 3, l'interpretazione grafica dei vincoli ESC: le diverse curve rappresentano i tracciati di f0, f1 ed f2, mentre i rettangoli colorati indicano le aree di variabilità imposte dal sistema. I rettangoli sono centrati verticalmente rispetto alle medie μ_{il} ed hanno altezza $2\gamma\sigma_i$ (ove possibile le medie sono state indicate con una linea orizzontale tratteggiata): appare chiaro che, a parità di campioni di riferimento del set ***Z***, modificando γ è possibile restringere o ampliare l'area di tolleranza, accettando un numero rispettivamente minore o maggiore di fonemi. La lunghezza delle frecce, presenti nella figura in basso, indica il valore $|\Delta f|_{il,max}$ per f1 ed f2.

Compiuta l'ultima analisi, il selettore è in grado di definire un sottogruppo ***W*** di fonemi sonori, sufficientemente stabili nel tempo da poter essere considerati candidate vocali accentate. Per ogni elemento $\boldsymbol{w_k}$ dell'insieme ***W***, con *k* minore o uguale a *l*, vengono memorizzati i valori medi μ_{1k} e μ_{2k}, perciò ogni picco è univocamente individuabile su un piano formantico (f1, f2). Termina in questa fase la prima parte della metodologia proposta.

L'insieme ***W*** sarà d'ora in avanti considerato esclusivamente come un insieme di punti (μ_{1k}, μ_{2k}) giacenti sul piano formantico. Tale insieme di punti sarà di seguito esplorato tramite algoritmi di ottimizzazione con il duplice scopo di classificare le diverse vocali accentate e di scartare gli elementi spuri, comunque presenti nell'insieme. Una rappresentazione grafica dell'insieme di punti ***W*** è riportata in figura 4 sinistra.

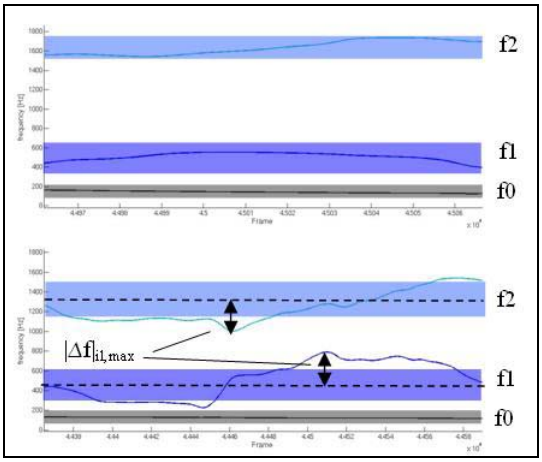

Figura 3: effetti del selettore ESC. Il fonema in alto viene accettato in quanto sufficientemente stabile, quello in basso viene scartato per l'eccessiva instabilità.

3. STIMA DEL TRIANGOLO VOCALICO

Durante le fasi di ricerca, i dati dell'insieme ***W*** sono stati confrontati con la segmentazione manuale, riscontrando un elemento comune in tutti i parlanti: il triangolo definito dai punti $m_k=(\mu_{1k}, \mu_{2k})$ risultava tendenzialmente sovrapposto con il triangolo vocalico dei diversi soggetti. In particolare le vocali risiedevano costantemente sul bordo di tale triangolo, mentre gli altri fonemi rimanevano all'interno di esso, come indicato in figura 4 centro. A partire da questa osservazione è stato messo a punto un sistema che stimasse le rette di tendenza del triangolo vocalico di ogni parlante. Per raggiungere tale scopo si è deciso di raggruppare i fonemi in tre grandi *cluster* (vedi figura 4 destra): un primo *cluster **A*** associato alla stima della retta delle vocali anteriori (retta *a*), un secondo *cluster **P*** associato alla stima della retta delle vocali posteriori (retta *p*) ed infine un terzo *cluster **C*** di fonemi centralizzati, associati alla bisettrice dell'angolo acuto β formato da *a* e *p* (retta *c*). Dal momento che la retta *c*, per costruzione, è vincolata alla posizione dalle altre due, ci si soffermerà, nel seguito, sul metodo impiegato per definire le rette del triangolo vocalico.

Le rette vengono definite in automatico, tramite un metodo autoassociativo ottenuto modificando il ben noto algoritmo k-means (di cui si parlerà nel dettaglio nel prossimo capitolo – dove verrà impiegato). A titolo introduttivo, si definisce con autoassociativo: un metodo iterativo in grado di adattarsi automaticamente ai dati, con lo scopo di raggrupparli secondo regole e vincoli imposti dall'utente, impiegando un numero di passi finito.

Tecnicamente, le regole prendono il nome di funzione obiettivo e il numero di passi è definito secondo i così detti criteri di arresto. Il metodo viene implementato in forma di problema di ottimizzazione, così che l'algoritmo si riconduca alla minimizzazione o massimizzazione vincolata (a seconda dell'esigenza) della data funzione obiettivo.

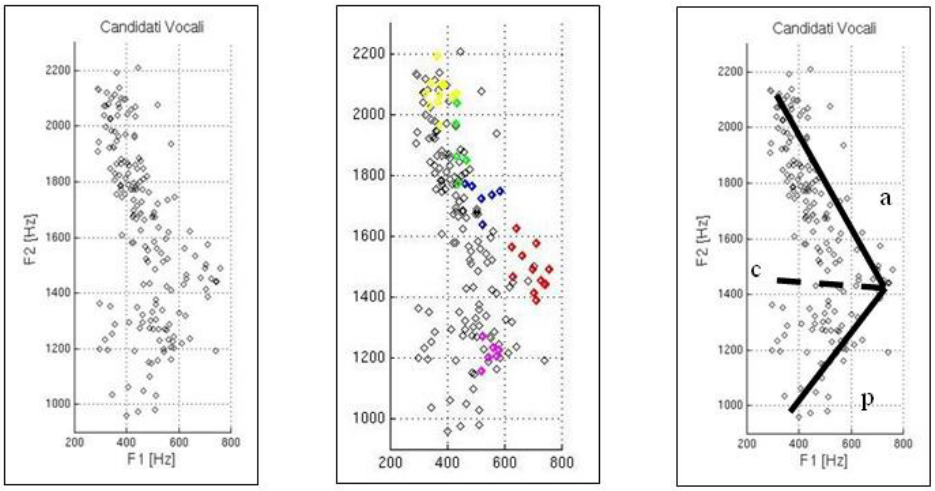

Figura 4: a sinistra rappresentazione grafica dell'insieme ***W***, al centro l'insieme sovrapposto alle vocali del parlante, a destra grafica delle rette *a*, *p* e *c* (tratteggiata).

Nello specifico, questo significa che l'algoritmo adottato deve essere in grado di creare tre gruppi di fonemi, tali per cui sia minimizzata la somma *S* delle distanze dei singoli punti appartenenti ad ***A*** e ***P***, rispetto alle rette *a* e *p*. In maniera compatta, il problema può essere formalizzato secondo la seguente scrittura:

(4) $$\arg\min_{\alpha_1,\alpha_2\in\Re^2} S = S(\alpha_1,\alpha_2) = \sum_{i=1}^{2}\sum_{t=1}^{n_i}\frac{\left|\mu_{2t} - \alpha_{i1}\mu_{1t} - \alpha_{i2}\right|}{\sqrt{1+\alpha_{i1}^2}}$$

Con:

$$\alpha_1 = \begin{bmatrix}\alpha_{11} & \alpha_{12}\end{bmatrix}$$
$$\alpha_2 = \begin{bmatrix}\alpha_{21} & \alpha_{22}\end{bmatrix}$$

Dove α_1 e α_2 rappresentano i coefficienti delle rette *a* e *p*, mentre n_i indica il numero di punti appartenenti rispettivamente ad ***A*** e ***P***. La funzione obiettivo non è altro che l'equazione della distanza euclidea di un punto da una retta.

Le rette *a* e *p*, vengono ricalcolate ad ogni passo, imponendo un classico problema ai minimi quadrati: rappresentano, dunque, le linee di tendenza degli ipotetici gruppi anteriore e posteriore. Definita con ΔS la differenza tra due successive catalogazioni, il criterio di arresto prevede che l'iterazione termini quando ΔS è nullo.

Il problema può essere spiegato in maniera intuitiva in questo senso: viene tentato un raggruppamento con tre rette "di primo tentativo". Ogni punto, rappresentante un fonema, viene associato alla retta più vicina. In questo modo tutti i punti sono inseriti in uno dei tre *cluster*. I punti del *cluster* ***C*** sono messi da parte (sono foni fortemente centralizzati) e viene calcolata la distanza di ogni punto dei *cluster* ***A*** e ***P*** dalle rette associate. Per entrambi i *cluster* viene calcolata la somma di tutte le distanze, ottenendo due numeri reali positivi.

Infine viene calcolata la somma di questi due valori, ottenendo un unico dato complessivo, indicato con *S*.

A questo punto l'algoritmo sostituisce le rette di primo tentativo con le rette di tendenza dei punti inseriti nei *cluster* ***A*** e ***P***. Essendo mutate *a* e *p*, viene ricalcolata di conseguenza anche la retta *c*. Ottenute delle nuove rette, diverse dalle prime, l'algoritmo annulla il precedente raggruppamento e ne realizza uno nuovo, che darà risultati in genere diversi dal primo: saranno ottenuti dei nuovi cluster ***A*** e ***P*** e muterà di conseguenza anche *S*. Confrontando la nuova classificazione con la vecchia, è possibile ottenere la variazione ΔS.

L'algoritmo prosegue con nuove iterazioni, fino a che è possibile osservare una stabilizzazione dei *cluster*. Di fatto, all'aumentare del numero di iterazioni diminuisce la differenza ΔS tra due catalogazioni successive, sino ad arrivare ad un punto in cui l'algoritmo non adduce alcun cambiamento ad una data configurazione. Questo stato rappresenta il così detto criterio di arresto dell'algoritmo.

Una volta terminato il processo, si è in grado di analizzare le diverse configurazioni ottenute. Dal momento che interessa la configurazione associata al valore minimo di *S*, indicato con S^* (vedi figura 5 destra), si considera le rette ottimali a^* e p^* ad esso associate, come la miglior stima possibile del triangolo vocalico del parlante (un esempio grafico è riportato in figura 5 centro).

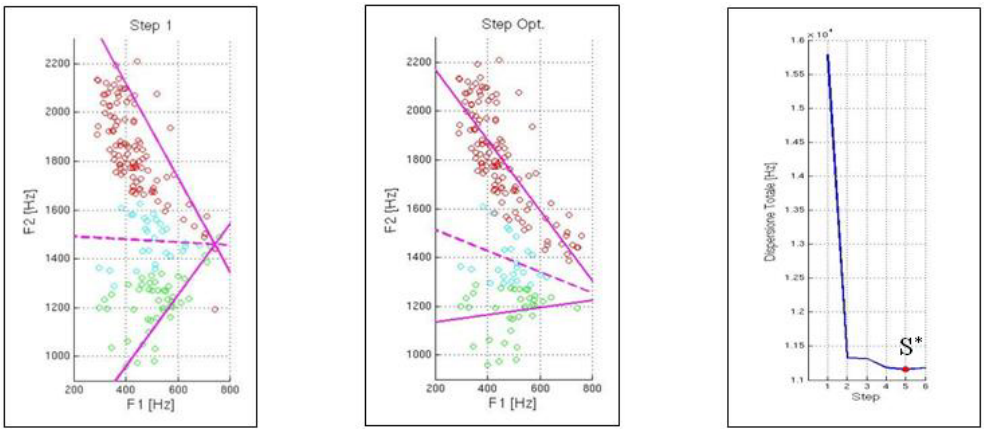

Figura 5: i cluster ***A*** (rosso), ***P*** (verde) e ***C*** (ciano) e la curva *S*. A sinistra: primo tentativo; al centro: configurazione ottima; a destra: indicazione di S^*.

Per quanto riguarda le rette di primo tentativo, esse sono definite a partire da tre punti generici ubicati ai presunti estremi del triangolo. In riferimento alle convenzioni adottate per la rappresentazione del piano formantico, vengono scelte le rette passanti: per il punto posto all'estremo in alto a sinistra del set di dati, per il punto posto all'estremo in basso a sinistra del set, per il punto più a destra in prossimità della frequenza f2 associata al baricentro del set (vedi figura 5 sinistra).

In ambito strettamente matematico, occorre rilevare che la convergenza dell'algoritmo di impiegato per l'individuazione del triangolo vocalico non è ancora stata propriamente provata. Conseguentemente, un ulteriore elemento da inserire nell'ambito dello studio riguarda proprio la dimostrazione di convergenza, che in questa fase è sempre stata raggiunta ma non assicurata a priori.

4. STIMA DEI CENTRI VOCALICI E SELEZIONE DEI CAMPIONI

Ottenuta la stima del triangolo vocalico, questa può essere impiegata per approssimare la posizione dei centri vocalici del parlante. Con centri vocalici, si intende qui la media dei valori di (μ_{1k}, μ_{2k}) associati ai campioni di una certa vocale. In maniera più formale, dato un generico *cluster* $\boldsymbol{R_i}$ di n_i punti P_t=(μ_{1t}, μ_{2t}), con $t=1,\ldots,n_i$, si definisce centro vocalico v_j, con j=1,..,7 per l'italiano, il punto di coordinate:

$$v_j = (v_{1j}, v_{2j}) = \left(\frac{1}{(d-s+1)} \sum_{j=s}^{d} \mu_{1j} \ , \frac{1}{(d-s+1)} \sum_{j=s}^{d} \mu_{2j} \right) \tag{5}$$

Con $1 \leq s \leq d \leq n_i$. Nel nostro caso i *cluster* cui si fa riferimento sono quelli delle vocali anteriori e posteriori: in altre parole, si calcolano i centri vocalici suddividendo i punti di ogni *cluster* in ulteriori sottoinsiemi, ciascuno associato ad una delle vocali dell'italiano.

A tal fine è possibile impostare un ragionamento analogo a quello proposto per le rette del triangolo vocalico: partendo da dei centri di tentativo, si cerca di definire i reali centri vocalici tramite iterazioni successive. La metodologia proposta è una basata su una versione leggermente modificata del problema noto come k-means, definito da J.B. MacQueen nel 1967. Il problema, si prefigge lo scopo di raggruppare un set di dati in un numero n prefissato di *cluster*, definito dall'utente. Tra i possibili *cluster* con cui si può raggruppare il set di dati, il problema vuol definire una configurazione ottimale, caratterizzata da particolari proprietà di compattezza.

Definiti con c_j i baricentri dei cluster, con j=1,…,n, ed indicati con P_i=(μ_{1i}, μ_{2i}) i punti appartenenti ad ogni singolo *cluster* $\boldsymbol{C_j}$, il problema può essere definito secondo la scrittura:

$$\underset{c_j \in \Re^2}{\arg\min}\ S = S(c_1,\ldots,c_n) = \sum_{j=1}^{n} \sum_{P_i \in C_j} \left\| P_i - c_j \right\| \tag{6}$$

Dove c_j, in accordo con l'equazione (5), rappresenta la media delle coordinate dei singoli punti P_i appartenenti al cluster $\boldsymbol{C_j}$. Nel caso in questione, una volta raggiunta la convergenza, i baricentri c_j rappresentano una stima ottima dei centri vocalici v_j. Il criterio di arresto è nuovamente dato dall'annullamento di ΔS tra due catalogazioni successive.

Uno dei più comuni metodi di risoluzione del problema, noto come algoritmo di Lloyd, prevede la definizione di un set di n baricentri di tentativo (spesso indicati con il termine "semi"), attorno ai quali aggregare i vari dati in gruppi, come descritto in Huang *et alii* (2001). L'algoritmo si sviluppa esattamente come descritto in precedenza per l'individuazione del triangolo vocalico (in effetti, l'algoritmo di Lloyd è il metodo da cui si è partiti per individuare il criterio di stima delle rette a e p). Dunque, si esegue una prima iterazione (vedi figura 6 sinistra), associando ogni elemento del *cluster* $\boldsymbol{A}$ ad un possibile baricentro delle vocali anteriori ed associando ogni elemento del *cluster* $\boldsymbol{P}$ ad un possibile baricentro delle vocali posteriori. I dati del *cluster* $\boldsymbol{C}$ sono considerati spuri ed esclusi a priori, data l'evidente centralizzazione. Conclusa la prima catalogazione, l'algoritmo calcolerà i baricentri dei *cluster* vocalici, ottenendo sette punti distinti. In seguito valuterà la somma S di tutte le distanze tra punti e baricentri ad essi associati. Eseguita questa operazione, la catalogazione viene annullata. I nuovi baricentri sono impiegati come nuovi semi, perciò verrà compiuta una nuova catalogazione e verrà computata un'ulteriore stima

di *S*. Anche in questo caso, catalogazioni diverse saranno associate a valori diversi di *S*, tuttavia tali differenze si assottiglieranno con l'aumentare delle iterazioni, sino ad annullarsi. In questo caso è possibile dimostrare che la configurazione più compatta è sempre quella ottenuta con l'ultima iterazione (vedi figura 6 destra). Dunque, terminate le successive catalogazioni, l'algoritmo è in grado di stimare i centri vocalici del parlante.

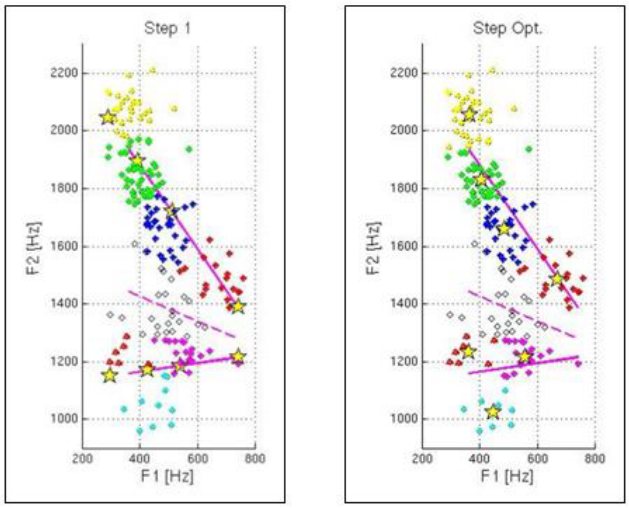

Figura 6: stima dei centri vocalici. A sinistra catalogazione iniziale, a destra catalogazione ottima (le stelle indicano le posizioni dei centri vocalici).

Le modifiche apportate al sistema originale di Lloyd sono due. In primo luogo esiste un problema relativo alla configurazione dei *cluster* vocalici: la vocale 'a' si presenta sia nel *cluster* ***A*** sia nel *cluster* ***P*** (in effetti, la sua presenza è possibile anche nel *cluster* ***C***, al momento trascurato), per tale motivo vengono definiti 8 semi, assegnando alla vocale 'a' due diversi *cluster* vocalici tra loro indipendenti – uno per il ramo anteriore ed uno per il posteriore. Terminate le iterazioni, i due *cluster* vengono uniti ed il centro vocalico viene stimato a partire dalla fusione dei dati, impiegando la (5). Tale soluzione non è probabilmente la più corretta formalmente ma ha dimostrato di essere efficace nei risultati.

In secondo luogo, esiste la possibilità che una vocale non sia presente tra i dati ricavati sino a questo punto (senza entrare nel merito dei sistemi non standard quali alcuni pentavocalici, si pensi all'effettiva assenza nel parlato di particolari vocali accentate). In questo caso è stata modificata la procedura di *bootstrap* dell'algoritmo. In pratica, dopo la prima catalogazione viene valutata l'esistenza o meno di dati all'interno dei *cluster* vocalici: se alcuni *cluster* rimangono vuoti alla prima iterazione, i semi ad essi associati sono eliminati e il primo passo viene ripetuto nuovamente con i semi rimanenti. È bene precisare che l'assenza di dati nella prima iterazione non è condizione né necessaria né sufficiente per dimostrare l'effettiva assenza di determinate vocali, tuttavia le prove sperimentali hanno indicato che l'errore indotto da questa semplificazione non è rilevante (quantomeno in relazione a questa fase di avanzamento del progetto, dove, si indicherà più avanti, esistono limitazioni più importanti).

Le condizioni iniziali per il posizionamento dei semi attingono nuovamente al lavoro di Cosi *et alii* (1995). All'interno dello studio sono riportate le statistiche relative alle vocali presenti nei bisillabi di AIDA. Tra i vari dati di interesse, sono presenti i valori medi di f1

ed f2 sia per gli uomini che per le donne. Tali valori possono essere assunti come i centri vocalici v_{mj} di un parlante italiano medio. A tali dati sarà dunque associato il triangolo vocalico medio dell'italiano (o una stima verosimile dello stesso). È possibile elaborare questi punti per ricavarne i semi utili per il *bootstrap* dell'algoritmo di catalogazione.

Innanzi tutto, i centri v_{mj} devono essere traslati in maniera tale da far coincidere il centro vocalico medio della 'a' con il punto di intersezione delle rette a^*, p^* e c^* di ogni parlante. Di seguito il ramo anteriore e posteriore del triangolo medio vengono scalati linearmente e indipendentemente sia lungo la direzione di f1 sia lungo f2. Le operazioni di traslazione e scalatura non fanno altro che spostare e deformare il triangolo del parlante medio sino a sovrapporlo con quello di un parlante generico. Compiute queste operazioni, i centri individuati da Cosi *et alii* (1995), si trovano giacenti sul triangolo del parlante oggetto di studio. Ovviamente, il posizionamento di tali punti non rappresenta l'individuazione di veri centri vocalici (basti pensare alle distanze relative tra le vocali), tuttavia è un primo passo per inizializzare la ricerca dei veri centri vocalici.

Una rappresentazione grafica di riferimento è riportata in figura 7: a sinistra sono indicati il triangolo vocalico di un parlante ed i centri vocalici medi dell'italiano (cerchi blu), al centro è visibile la loro proiezione sul triangolo del parlante (stelle).

Una volta concluse le operazioni di stima ottima dei centri vocalici, i *cluster* associati a tali centri raccolgono tutti i foni appartenenti al ramo anteriore e posteriore del triangolo vocalico. In tale configurazione sono catalogati sia gli effettivi fonemi vocalici accentati sia una parte di dati spuri, relativamente distanti dai centri c_j. Al fine di affinare la catalogazione ed eliminare i dati spuri, viene imposta una soglia di accettazione regolabile dall'utente. Tale soglia è stimata in maniera dinamica al variare della covarianza dei *cluster* vocalici.

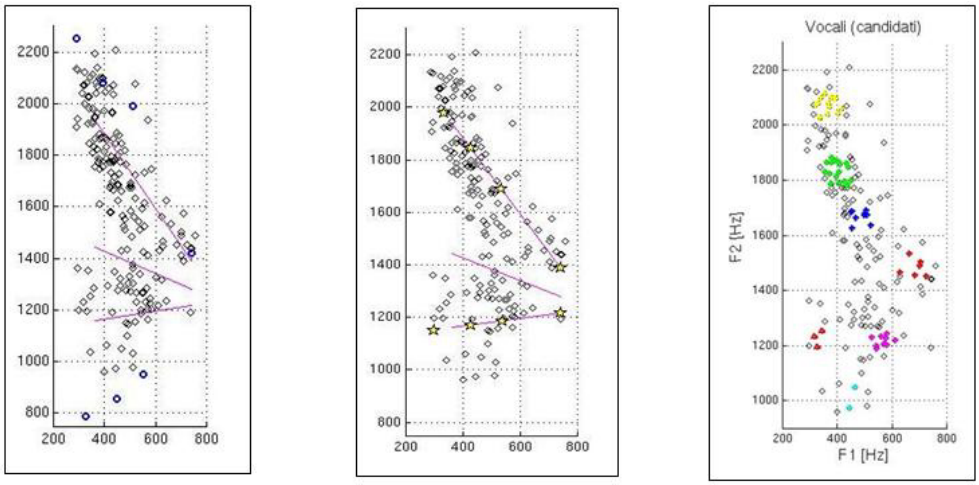

Figura 7: adattamento dei centri vocalici medi (Cosi *et alii*, 1995) al triangolo vocalico di un parlante casuale (sinistra e centro) e presunti fonemi vocalici dopo la sogliatura imposta dall'operatore (destra).

In maniera più precisa, si è assunto in questa sede che i *cluster* vocalici fossero ben modellabili con distribuzioni gaussiane multivariate a 2 dimensioni. Le medie stimate di tali distribuzioni coincidono con i centri c_j dei *cluster*, mentre le matrici di covarianza Σ_j

sono stimate a partire dai dati (spuri o meno) presenti nei *cluster* stessi. Tali parametri permettono di individuare un'ellisse sul piano formantico in prossimità di ogni *cluster*: i centri vocalici coincidono con i centri delle ellissi, mentre gli assi delle stesse e la loro rotazione sono definiti in base a Σ_j.

La soglia di filtraggio è regolata in funzione della matrice Σ_j: l'utente indica un coefficiente positivo ξ con cui scalare le dimensioni delle ellissi associate ai *cluster*. Valori di ξ maggiori di 1 aumentano l'area dell'ellisse, mentre valori minori di 1 la riducono. Il sistema di filtraggio trattiene solo i punti interni all'ellisse di ogni vocale, scartando i dati esterni. In questo modo è possibile limitare la selezione dei fonemi, trattenendo solo i presunti fonemi vocalici accentati (figura 7 destra).

5. SPERIMENTAZIONE

Una volta ipotizzato un possibile metodo di ricerca delle vocali accentate, il primo passo della sperimentazione è stato quello di definire un parlato di natura generica ma ripetibile, che fosse noto a priori durante i test. A tale scopo ci si è orientati sul parlato letto. Il testo scelto è stato liberamente tratto da Georges (1994). Data la brevità, la porzione impiegata è riportata di seguito in maniera integrale:

"I primi esseri viventi hanno cominciato a esprimersi a gesti che, per chi stava loro accanto, sono diventati segni, tipici del linguaggio orale. I segni della gestualità precisano la parola, la enfatizzano e spesso ne prendono il posto. Il corpo parla al di là della parola e della scrittura; consente di esprimere i propri sentimenti e di comunicare con i propri simili anche senza menzionare alcuna parola. I segni del corpo sono la forma prima del linguaggio. I neonati, ben prima di aver imparato a parlare, si esprimono con il corpo e con diverse manifestazioni fonetiche: strilli, gemiti e risa. Per chi li deve interpretare questi segnali sono solo indizi. A poco a poco però il bambino imparerà non solo il sistema fonetico della lingua madre ma anche il linguaggio somatico, cioè del corpo."

5.1 Costituzione della base dati

Sono stati acquisiti 16 parlanti maschili di età compresa tra i diciannove ed i trenta anni, principalmente studenti dei corsi di laurea in ingegneria tenuti presso l'Ateneo. Nove di essi sono stati acquisiti per tre volte, a distanza di circa dieci giorni tra una sessione e l'altra; tre di essi sono stati acquisiti per due volte, secondo le medesime modalità; i rimanenti parlanti sono stati acquisiti una sola volta. In totale sono state acquisite 37 tracce. Ad ogni soggetto è stata fatta raccomandazione di leggere in maniera spontanea, senza preoccuparsi dei possibili difetti di dizione, trascurando gli sbagli e, qual'ora si sia presentata qualche manifestazione spontanea estranea al testo (interruzione per risate, per perdita del segno nel testo e quant'altro), è stato chiesto di non interrompere la sessione, riprendendo la lettura dal punto che si ritenesse più opportuno. Nel caso in cui dette manifestazioni abbiano permesso di acquisire enunciati estranei al testo (eventi piuttosto rari), tali enunciati sono stati esclusi nelle fasi di elaborazione.

Le acquisizioni si sono tenute in un locale dell'università adiacente alle aule presso cui si tenevano le lezioni, non dunque in cabine appositamente preparate. Di conseguenza nel locale, arredato a sala conferenze, era presente del riverbero. Inoltre, pur limitando gli eventi accessori, le registrazioni contengono altri segnali oltre a quello vocale. Principalmente si tratta di eventi generati dal lettore stesso, come ad esempio la respirazione o lo stropiccio dei fogli di lettura contro il microfono.

La catena di acquisizione è stata realizzata impiegando un microfono a doppio condensatore, normalmente impiegato per registrazioni musicali. La scelta del doppio condensatore è stata dettata dalla possibilità di disattivare uno dei due elementi, al fine di limitare l'area di acquisizione nella sola direzione del parlante (la sensibilità del microfono era ridotta nelle altre direzioni). Il microfono, preamplificato, trasmetteva il proprio segnale ad una scheda audio munita di porta usb, che scaricava i dati direttamente sul pc. Al fine di mantenere un livello di segnale relativamente costante fra le varie tracce, si è agito sul guadagno della strumentazione per compensare la fonazione dei diversi parlanti. Per la registrazione dei dati è stato utilizzato il software *audacity*, rilasciato su licenza GPL. Il formato wave è stato scelto per la memorizzazione dei dati. Il segnale è stato codificato in PCM mono con frequenza di campionamento pari a 44100Hz e quantizzazione a 16 bit. La codifica scelta è in realtà ridondante rispetto alle effettive necessità del progetto, tuttavia si è preferito acquisire i dati in alta qualità, avendo comunque la possibilità di alterare la stessa in un secondo tempo (ad esempio per produrre ulteriori test con codifica telefonica).

Data la necessità di verificare le capacità discriminatorie e classificatorie del sistema, le tracce sono state innanzi tutto elaborate manualmente tramite *Praat*, isolando i fonemi vocalici di interesse.

5.2 Analisi dei risultati

Al fine di valutare l'efficacia dei diversi passaggi costituenti il metodo proposto, si riportano di seguito alcuni dati relativi ai risultati di classificazione. I dati fanno riferimento a due diverse configurazioni: la prima è priva di soglia di filtraggio finale ($\xi=+\infty$), la seconda è stata ottenuta imponendo una soglia di filtraggio $\xi=+2$. La seconda scelta è dovuta al fatto che tale soglia ha rappresentato con buona approssimazione la configurazione di equal error rate (EER). Per il calcolo dell'EER si è fatto riferimento alla percentuale di dati spuri trattenuti dal sistema ed a quella di vocali accentate respinte; entrambe le percentuali sono state calcolate al termine della catena di selezione, catalogazione e filtraggio.

La tabella sottostante riporta il numero totale di vocali accentate campionate manualmente da tutte le tracce:

a	ɛ	e	i	ɔ	o	u
336	222	143	280	203	135	0

Innanzi tutto si è analizzato l'efficacia del selettore IBT-DESF ed il contributo prodotto dal metodo di filtraggio posto in coda alla procedura (tramite la variabile ξ). La loro resa è ricavabile dall'osservazione degli errori di Tipo I e II. Relativamente alle cause di tali errori, si ricorda che tutte le condizioni imposte dal selettore IBT-DESF sono strettamente necessarie ma non sufficienti, per tale motivo è congenita la presenza di errori di tipo I (falsa accettazione). Parimenti, data la natura delle soglie di filtraggio, risulta inevitabile anche un errore di tipo II (falsa reiezione). Infine c'è da precisare che la costante γ, così come è stata calcolata, causa un limitato errore di tipo II. Le percentuali di tali errori, nelle due configurazioni, sono risultate pari a:

	Tipo I	Tipo II
$\xi=+\infty$	62.5	5.9
$\xi=+2$	39.7	39.0

La tabella indica che la capacità di selezione del metodo IBT-DESF non è particolarmente performante: una gran parte dei foni spuri continua a permanere al momento di individuare triangolo e centri vocalici. Al fine di raggiungere una configurazione di equilibrio (EER) si è costretti ad intervenire con un filtraggio molto selettivo, causando l'innalzamento del punto di EER. Scendendo nel dettaglio, le percentuali di rumore residuo (foni spuri) e di vocali non catalogate sono così ripartire:

		a	ɛ	e	i	ɔ	o	u
$\xi=+\infty$	Tipo I	9.0	13.8	17.6	10.9	5.2	3.5	2.4
	Tipo II	7.4	0.5	0	0	16.8	17.5	-
$\xi=+2$	Tipo I	5.4	8.9	11.3	7.1	3.1	2.2	1.6
	Tipo II	38.3	57.7	35.0	21.8	42.4	48.2	-

Relativamente alla capacità discriminatoria del sistema, ovvero alla capacità di individuare il triangolo ed i centri vocalici, sono state calcolate le matrici di confusione nelle due configurazioni. I valori sono espressi in riferimento al numero di vocali campionate manualmente, di conseguenza, sono affetti dall'errore di tipo II indotto sia dal selettore sia dalla soglia. Si ricorda che le colonne indicano il tipo di vocale, mentre le righe indicano la classificazione eseguita dal sistema. Matrice di confusione nel caso $\xi=+\infty$:

	a	ɛ	e	i	ɔ	o	u
a	***89.3***	1.4	0	0	6.4	5.9	-
ɛ	3.3	***71.6***	*5.6*	0	0	0	-
e	0	*25.7*	***68.5***	6.4	0	0	-
i	0	0.9	25.9	***93.6***	0	0	-
ɔ	0	0	0	0	***50.7***	*17.8*	-
o	0	0	0	0	*21.7*	***41.5***	-
u	0	0	0	0	4.4	13.3	-

Matrice di confusione nel caso $\xi=+2$:

	a	ɛ	e	i	ɔ	o	u
a	***61.6***	0.5	0	0	3.5	2.4	-
ɛ	0	***37.4***	*3.5*	0	0	0	-
e	0	*4.5*	***53.1***	3.6	0	0	-
i	0	0	8.4	***74.6***	0	0	-
ɔ	0	0	0	0	***37.0***	*9.4*	-
o	0	0	0	0	*13.8*	***30.6***	-
u	0	0	0	0	3.5	9.4	-

Le matrici di confusione indicano che le vocali vengono in buona parte riconosciute (le percentuali sulla diagonale sono molto superiori a quelle fuori diagonale), se ne deduce, quindi, che i sistemi di catalogazione riescono comunque ad effettuare delle buone stime

del triangolo e dei centri vocalici. Si tenga conto che i parametri fuori diagonale, relativi alle coppie ɛ/e - ɔ/o, sono comunque affetti da un inevitabile errore, dovuto alla necessità di porre un confine netto tra foni, che spesso sfumano in maniera continua da un dominio all'altro. Per tale motivo sono stati indicati in corsivo.

Le conclusioni che si possono dedurre dai due esperimenti pilota, indicano che i principali limiti della tecnologia riguardano il selettore IBT-DESF. In ogni caso si è dimostrato poco capace di filtrare i foni spuri, specie quelli relativi alle vocali posteriori. In molti casi la stima del triangolo vocalico tendeva a evidenziare un ramo posteriore più centralizzato di quello reale, con ovvie imprecisioni che si ripercuotevano sull'individuazione dei centri vocalici. Uno degli effetti più evidenti di questo fenomeno è l'impropria individuazione di u. Come si evince anche dalla figura 6 destra, la presenza di foni spuri provoca spesso un impreciso posizionamento dell'ipotetico centro vocalico delle u, che si trova addirittura a valori di f1 minori del centro vocalico della i. Altre volte, la presenza di foni spuri si ripercuote addirittura sull'individuazione di tutte le vocali posteriori, la cui catalogazione risulta leggermente traslata rispetto alla reale posizione. Al contrario, le vocali anteriori risultano poco influenzate dai foni spuri. È altresì vero che il selettore risulta relativamente più efficace sul ramo anteriore, riducendo i fenomeni di rumorosità. Una delle possibili cause della differente capacità di filtraggio può essere associata ad una carente od erronea raccolta di informazioni riguardo alla stabilità delle vocali posteriori, mentre la carenza generale di selettività potrebbe essere parzialmente compensata con regolazioni migliori dei coefficienti empirici.

Le conseguenze sulla fase di catalogazione dei foni sono ovvie: data l'impossibilità di definire la differenza tra un fono accentato o meno, a partire dalla mera posizione sul piano f1-f2, la classificazione finale conduce a trattenere alcuni foni erronei, escludendone altri invece rilevanti. Risulta importante, quindi, migliorare il selettore.

Come nota a margine, si fa notare che la segmentazione manuale di confronto, eseguita da personale non specializzato, è sì un primo utile termine di paragone, ma è risultata anch'essa non priva di imperfezioni. Questo ci spinge a ritenere che alcuni dati delle matrici di confusione potrebbero risultare qualitativamente migliori, se paragonati con una segmentazione più accurata.

6. CONCLUSIONI E SVILUPPI FUTURI

Il presente articolo descrive la struttura e la prima sperimentazione di un sistema specializzato per l'individuazione e la catalogazione di fonemi vocalici. Il metodo è stato ipotizzato nell'ambito di un progetto di ricerca relativo al riconoscimento automatico del parlante, in essere presso il DMTI. Uno degli obiettivi per lo sviluppo di tale sistema è stato quello di realizzare uno strumento totalmente automatico, che richiedesse costi di taratura minimi. Allo stato attuale della ricerca è stato sviluppato un prototipo funzionante di tale sistema. Il prototipo, nella configurazione attuale, è stato testato solo in condizioni testo-dipendenti su parlato letto in laboratorio.

La tecnologia impiegata è relativamente semplice e si basa su principi matematici elementari come le rette di regressione (analisi dei dati ai minimi quadrati) e la catalogazione tramite k-means (algoritmo di classificazione di tipo deterministico). Più in specifico, la procedura di classificazione si appoggia su parametri quali l'intensità della traccia audio, la frequenza fondamentale e le prime due formanti. Tale strumento ha lo scopo di eseguire una prima selezione grossolana, che consenta di isolare segmenti classificabili come foni sonori.

Gli elementi così isolati, rappresentati su un piano formantico F1-F2, appaiono racchiusi in un triangolo, limitato, appunto, dalle vocali accentate. Tramite il posizionamento ottimale delle rette di regressione, è possibile suddividere iterativamente i foni in tre gruppi: posteriore, anteriore e dei foni centralizzati. Al termine dell'iterazione si ottiene una stima approssimata del triangolo vocalico sottostante i foni.

Infine, il sistema di riconoscimento prevede la classificazione delle sette vocali tramite il noto algoritmo k-means e la selezione dei fonemi più rappresentativi tramite l'analisi della covarianza dei singoli *cluster*. Quest'ultima fase è stata parametrizzata, così da regolare il dettaglio del sistema in base alle proprie esigenze. Tramite questa accortezza il prototipo potrà essere impiegato in maniera flessibile sia come strumento di screening che come catalogatore vero e proprio.

La sperimentazione pilota mostra che la classificazione prodotta, anche tenendo conto di un possibile affinamento, risulta sicuramente meno accurata di quella ottenibile con mezzi più sofisticati proposti dalla recente letteratura (vedi Huang *et alii* (2001)). Tuttavia, se talune ricerche richiedono un'analisi estremamente accurata dei dati, dall'altro lato esistono svariate altre applicazioni, specie nel campo della biometrica, per le quali lo strumento, al termine del suo sviluppo, risulterà adeguato. Inoltre, anche nelle applicazioni più esigenti, lo strumento può risultare un ottimo mezzo per lo screening dei corpora con cui un fonetista deve confrontarsi giornalmente.

Lo sviluppo previsto per il progetto è attualmente articolato in due fasi: la prima mira ad affinare la precisione di selezione, migliorando le caratteristiche del selettore IBT-DESF o, nel caso si dimostri necessario, valutando la possibilità di integrare tale selettore.

La seconda fase di sviluppo mira ad applicare il sistema a corpora di parlato spontaneo o semispontaneo acquisiti sia in laboratorio che telefonicamente. L'obiettivo finale, al momento ancora distante, è quello di implementare uno strumento in grado di individuare i fonemi di interesse su acquisizioni del mondo reale (ad esempio conversazioni telefoniche vere) così da facilitare ed automatizzare le procedure di preprocessamento di un sistema di riconoscimento vocale.

BIBLIOGRAFIA

Alfano, I. (2005), La percezione dell'accento lessicale: un test sull'italiano a confronto con lo spagnolo, in *Atti del II Convegno Nazionale AISV 2005*, Salerno, 30 Novembre – 2 Dicembre.

Cosi, P., Ferrero, F. & Vagges K. (1995), Rappresentazioni acustiche e uditive delle vocali italiane, in *Atti del XXIII Congresso Nazionale AIA 1995*, Bologna, 12-14 Settembre, pp. 151-156.

Georges, J.(1994), Il linguaggio dei segni, la scrittura e il suo doppio, Electa/Gallimard.

Huang, X., Acero, A. & Hon, H. (2001), *Spoken language processing: a guide to theory, algorithm and system development*, Washington: Prentice Hall.