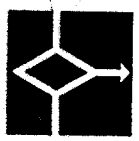

A.I.C.A.
Associazione Italiana per l'Informatica
ed il Calcolo Automatico

CONGRESSO ANNUALE
ANNUAL CONFERENCE

ATTI
PROCEEDINGS

Volume 1

BARI 19-21 SETTEMBRE 1990

Adattamento ad applicazioni diverse e sperimentazione di un riconoscitore della voce per grandi vocabolari

P. Alto, M. Brandetti, M. Ferretti, G. Maltese, F. Mancini, A. Mazza, S. Scarci, G. Vitillaro

IBM Italia Centro Ricerca di Roma
via Giorgione 159, 00147 ROMA

Sommario

Presso il Centro di Ricerca IBM di Roma e' stata sviluppata una macchina per dettare in grado di riconoscere in tempo reale frasi in linguaggio naturale. Il primo sistema realizzato e' costituito da un personal computer equipaggiato con quattro schede specializzate. Esso utilizza un vocabolario orientato ad un lessico economico e finanziario contenente piu' di 20000 parole. Gli esperimenti di uso della macchina, condotti presso il Centro IBM, hanno dimostrato la sua validita' come strumento per la creazione automatica dei testi. Terminata la prima fase di sperimentazione e' sorta la necessita' di studiare come il riconoscitore viene utilizzato durante la normale attivita' lavorativa. A questo scopo sono state prese in considerazione due applicazioni, una riguardante la dettatura di referti radiologici, l'altra la preparazione di polizze di assicurazione.
In questo articolo vengono descritte le tecniche messe a punto per permettere un rapido adattamento del riconoscitore al lessico della applicazione in esame; vengono inoltre riportati i risultati ottenuti nell'impiego del prototipo.

Introduzione

Negli ultimi anni lo sviluppo della tecnologia del riconoscimento della voce ha permesso la creazione di sistemi per il riconoscimento del linguaggio naturale su grandi vocabolari in tempo reale [1] [2]. Contemporaneamente l'aumento delle possibilita' di integrazione dei componenti offerte dalla microelettronica ha permesso di poter disporre di riconoscitori basati su personal computer. La disponibilita' di sistemi con queste caratteristiche ha reso possibile l'effettuazione di studi sui fattori umani legati all'introduzione di uno strumento cosi' innovativo. Presso il Centro di Ricerca IBM di Roma sono stati effettuati esperimenti per determinare l'accettazione da parte degli utenti della dettatura automatica [3]. Il vocabolario di 20000 parole utilizzato per questi esperimenti e' adatto alla dettatura di testi di economia e finanza. Nel seguito faremo riferimento a questa applicazione con la sigla EF.

I risultati ottenuti hanno dimostrato che il riconoscimento della voce puo' offrire una valida alternativa ai mezzi normalmente utilizzati per la creazione dei testi. Per rendere piu' significativi gli esperimenti di dettatura si e' pensato di utilizzare il prototipo di macchina per dettare in un ambiente di lavoro. Sono state selezionate due applicazioni che per le loro caratteristiche sono apparse adatte ai nostri scopi. La prima applicazione considerata concerne la dettatura di referti radiologici. Questa verra' indicata nel seguito con la sigla **RR**. La seconda applicazione riguarda la stesura di polizze di assicurazione e verra' indicata nel seguito con la sigla **PA**. Un problema che si e' posto per la realizzazione dei riconoscitori prototipo da utilizzare nei due esperimenti e' stato quello di adattare il nostro vocabolario base al lessico richiesto dalle diverse applicazioni. Cio' ha comportato lo sviluppo di tecniche *ad hoc* per permettere un rapido adattamento dei modelli acustico e linguistico del riconoscitore. I prototipi realizzati sono stati installati nell'ambiente di lavoro ed utilizzati da utenti senza nessuna precedente esperienza di dettatura automatica per lo svolgimento delle loro usuali attivita'.

Struttura del riconoscitore della voce

Il sistema da noi utilizzato e' basato sull'approccio statistico al riconoscimento della voce [1] [2][4]. In questo approccio si cerca di determinare la sequenza di parole $\hat{W}$ che ha la massima probabilita' data l'informazione acustica $\overline{A}$ estratta dal segnale che e' stato osservato. Nel nostro caso il segnale acustico viene codificato come una sequenza di **etichette acustiche** calcolate dal segnale ogni centesimo di secondo. Applicando il teorema di Bayes possiamo scrivere:

$$P(\overline{W}\,|\,\overline{A}) = \frac{P(\overline{A}\,|\,\overline{W})P(\overline{W})}{P(\overline{A})} \qquad (1)$$

dove $P(\overline{W}\,|\,\overline{A})$ e' la probabilita' che la sequenza di parole $\overline{W}$ produca l'informazione acustica $\overline{A}$. $P(\overline{W})$ e' la probabilita' a priori della sequenza di parole $\overline{W}$. $P(\overline{A})$ e' la probabilita' della sequenza di informazioni acustiche $\overline{A}$. Il nostro obiettivo e' quello di trovare il massimo della precedente espressione al variare di $\overline{W}$. $P(\overline{A})$ puo' essere trascurato dato che non dipende da $\overline{W}$. Il problema si riduce percio' a quello di trovare il massimo del numeratore della frazione.

Come conseguenza di queste considerazioni il problema del riconoscimento della voce puo' essere scomposto nei seguenti sottoproblemi:

1. **Elaborare il segnale vocale** per codificarlo nella stringa $\overline{A}$;
2. Calcolare la probabilita' $P(\overline{A}\,|\,\overline{W})$ (questo e' il compito del **modello acustico**);
3. Calcolare $P(\overline{W})$ (compito del **modello del linguaggio**);
4. Cercare mediante una efficiente **strategia di ricerca** la sequenza di parole che ha la massima probabilita'.

Mentre la fase di elaborazione acustica e la strategia di ricerca possono essere considerate indipendenti dall'applicazione, i modelli acustico e linguistico debbono essere modificati a seconda del vocabolario utilizzato. Nei prossimi paragrafi verranno illustrate le tecniche impiegate per effettuare l'adattamento dei due modelli.

Adattamento del modello acustico

Come abbiamo visto nel paragrafo precedente il compito del modello acustico nella struttura di un riconoscitore statistico della voce e' quello di calcolare la probabilita' $P(\overline{A}\,|\,\overline{W})$. Nell' approccio statistico il modello acustico e' costituito da *sorgenti di Markov*. Una sorgente di Markov e' una macchina probabilistica a stati finiti. Ad intervalli di tempo fissi la sorgente effettua una transizione che puo' provocare o meno il cambiamento di stato della macchina; contemporaneamente viene prodotta una etichetta acustica [2]. Sia le transizioni che l'emissione delle etichette acustiche avvengono in base a distribuzioni di probabilita' che dipendono soltanto dallo stato in cui la sorgente si trova e non dalla storia precedente. Mentre e' possibile osservare la sequenza di etichette prodotta, la sequenza di stati visitati dal modello rimane nascosta. Questi modelli vengono quindi chiamati *sorgenti di Markov nascoste*.

Nel modello acustico del riconoscitore ognuna delle parole appartenenti al vocabolario e' rappresentata da un modello costituito da una sorgente di Markov. Per la costruzione di questi modelli e' possibile utilizzare due tecniche diverse. Una tecnica prevede la costruzione automatica del modello di una parola utilizzando varie pronunce della parola stessa effettuate da diversi parlatori [5]. Un'altra tecnica e' quella di definire un alfabeto di unita' acustiche con le quali rappresentare i suoni base della lingua e costruire il modello per la parola mediante la concatenazione dei modelli di Markov che rappresentano le unita' acustiche. Esempi di unita' acustiche utilizzate per il riconoscimento della voce sono: sillabe, difoni, fonemi. Nel nostro caso e' stata utilizzata quest'ultima tecnica. Il tipo di unita' acustica scelto e' stato il fonema. Un alfabeto di 56 unita' fonetiche e' stato creato per descrivere i suoni elementari dell'italiano [4]. Per ogni unita' fonetica esiste una corrispondente sorgente di Markov che la

rappresenta. Per ogni parola del vocabolario la sorgente markoviana corrispondente viene costruita mediante la concatenazione delle sorgenti di Markov fonetiche che costituiscono la pronuncia della parola. Nel nostro sistema tutte le unita' fonetiche presentano la stessa struttura. La distinzione tra diverse unita' fonetiche e' lasciata interamente alle probabilita' di transizione da uno stato all'altro e dalla probabilita' di emissione delle etichette acustiche. La stima di queste probabilita' avviene nella fase di *addestramento acustico* durante la quale l'utente pronuncia un testo noto.

Il primo passo da compiere quando si vuole adattare il riconoscitore ad una nuova applicazione e', una volta definito il nuovo vocabolario, effettuare la trascrizione fonetica di tutte le parole. Allo scopo di limitare il numero di trascrizioni fonetiche da effettuare per una nuova applicazione e' stato costruito un database contenente tutte le parole e le relative trascrizioni fonetiche che hanno fatto parte di vocabolari utilizzati per altre applicazioni. Effettuando una ricerca nel database e' possibile individuare tutte le parole per le quali non esiste gia' la trascrizione fonetica. Di solito la trascrizione fonetica e' un processo che viene compiuto in modo completamente manuale risultando cosi' costoso e facilmente soggetto ad errori. Per questa ragione si e' cercato di realizzare degli strumenti automatici che rendessero tale processo piu' rapido ed affidabile. I sistemi proposti per la trascrizione fonetica automatica sono basati su regole [6] [7] o su apprendimento automatico da un insieme di dati di addestramento [8]. Questi sistemi non permettono pero' di raggiungere l'accuratezza desiderata per un riconoscitore della voce. Cio' e' dovuto alla intrinseca complessita' del problema ed alla difficolta' di descrivere, tramite un limitato insieme di regole, tutti i casi possibili. L'approccio seguito per risolvere il problema e' stato quello di separare la conoscenza fonotattica da quella lessicale. Infatti, mentre la prima puo' essere descritta da un limitato insieme di regole, la seconda e' basata in gran parte sull'esperienza umana e non puo' essere facilmente formalizzata [9]. Il sistema costruito e' in grado, data la stringa ortografica che rappresenta la parola, di produrre un insieme di possibili trascrizioni fonetiche che si ottengono applicando le regole per il passaggio da forma ortografica a forma fonetica. La scelta della trascrizione corretta, date tutte le alternative, avviene manualmente. Un insieme di 78 regole permette di descrivere le ambiguita' esistenti nell'italiano per la traduzione della parola dalla forma ortografica a quella fonetica. Ogni regola e' costituita da una parte sinistra ed una parte destra. La parte sinistra e' composta da una stringa ortografica, la parte destra e' costituita dalle possibili trascrizioni fonetiche della stringa. Le possibili trascrizioni alternative prodotte da questo insieme di regole vengono sottoposte ad una successiva analisi mediante un insieme di regole globali che eliminano le trascrizioni impossibili, ad esempio quelle che contengono piu' di una vocale accentata. In media il numero di trascrizioni prodotte per parola e' 5. La trascrizione fonetica corretta e' sempre presente nelle alternative che vengono presentate all'operatore. Quest'ultimo, sulla base della conoscenza lessicale e semantica della parola, e' in grado di selezionare facilmente la trascrizione corretta che verra' utilizzata per rappresentare la pronuncia della parola nel riconoscitore. Questa tecnica si e' dimostrata estremamente efficace ed ha permesso un rapido adattamento del modello acustico del riconoscitore della voce al nuovo vocabolario.

Test di laboratorio del riconoscitore

Prima di sperimentare il comportamento del riconoscitore nell'ambiente in cui dovra' operare, vengono solitamente effettuati dei test di riconoscimento in laboratorio. Il test viene realizzato facendo dettare un testo caratteristico dell'applicazione a diversi parlatori e calcolando il tasso di riconoscimento ottenuto. Per rendere l'esperimento significativo e' importante che il test contenga, in contesti tipici dell'applicazione, tutte le unita' fonetiche utilizzate per la costruzione del modello acustico. L'insieme di frasi da utilizzare per effettuare queste verifiche viene di solito costruito manualmente cercando di rappresentare con il minor numero di frasi il maggior numero possibile di contesti fonetici. E' stato realizzato un insieme di procedure per rendere il processo di creazione del test completamente automatico. Il programma utilizza come dati di partenza un insieme di frasi tipiche della applicazione. Normalmente vengono utilizzati i dati di addestramento per il modello del linguaggio. Su questo insieme di frasi viene effettuata un'analisi preliminare volta ad eliminare tutte quelle che contengono parole non presenti nel nuovo vocabolario. Come prima frase del test viene selezionata quella contenente il maggior numero di foni diversi. Il resto del test viene creato incrementalmente

aggiungendo ad ogni passo la frase che presenta il punteggio piu' alto dato l'insieme di quelle gia' selezionate. Il calcolo del punteggio avviene in questo modo:

- si considerano le frequenze di occorrenza di ogni unita' fonetica nell'insieme di frasi di test finora selezionato;
- per ogni frase contenuta nei dati a disposizione e non ancora selezionata si calcola un punteggio utilizzando la seguente formula:

$$P(S_k) = \sum_i f_i \exp(-h_i) \quad (2)$$

dove f_i e' la frequenza del fone i nella frase S_k e h_i e' la frequenza del fone i nelle frasi selezionate fino a quel momento;

- la sommatoria e' estesa soltanto ai foni presenti nelle frasi gia' selezionate meno di un numero di volte prefissato.

Applicando questo algoritmo e' possibile selezionare in modo efficiente ed automatico un insieme di frasi contenenti contesti fonetici tipici dell'applicazione ed in grado di essere un valido test del modello acustico costruito.

Adattamento del vocabolario.

Poiche' il riconoscitore e' basato su un vocabolario predefinito, la scelta del dizionario ne condiziona in modo essenziale l'utilita'.

Il vocabolario utilizzato per gli esperimenti finora effettuati (EF) e' orientato verso un dominio linguistico di tipo economico-finanziario. Le 20000 parole che lo costituiscono sono infatti le piu' frequenti riscontrate in un *corpus* di 44 milioni di parole composto di articoli di giornali (*Il Sole 24 Ore*), settimanali (*Il Mondo*) e comunicati dell'agenzia ANSA. Tale dizionario ha una copertura del 96.5% su testi di economia e finanza non utilizzati per la sua creazione.
Questo tipo di lessico e' pero' molto "distante" da quello impiegato per la compilazione di referti radiologici, mentre mostra una maggiore affinita' con i testi prodotti dalla compagnia assicuratrice. Nel primo caso si e' quindi ritenuto necessario ricostruire interamente il dizionario basandosi esclusivamente su un corpus di referti radiologici, nel secondo caso si e' costruito un dizionario misto , basato cioe' su EF e arricchito con le parole piu' frequenti del corpus applicativo.

Di seguito viene descritto con maggiore dettaglio il processo che ha portato alla definizione dei vocabolari per i due esperimenti focalizzando l'attenzione sulle principali caratteristiche dei corpora che sono stati analizzati e sui criteri utilizzati per la selezione delle parole.

Applicazione RR Per la definizione del vocabolario costruito per la dettatura dei referti radiologici, e' stato analizzato un corpus di 5 milioni di parole (OS) costituito da referti raccolti presso 4 ospedali (OS1, OS2, OS3, OS4), e un corpus di 50000 parole (VI) raccolto presso il reparto di radiologia che ha ospitato l'esperimento.
Per la scelta della dimensione del vocabolario ci siamo basati sull'analisi dell'andamento della copertura del testo in funzione del numero di parole del dizionario (figura 1).

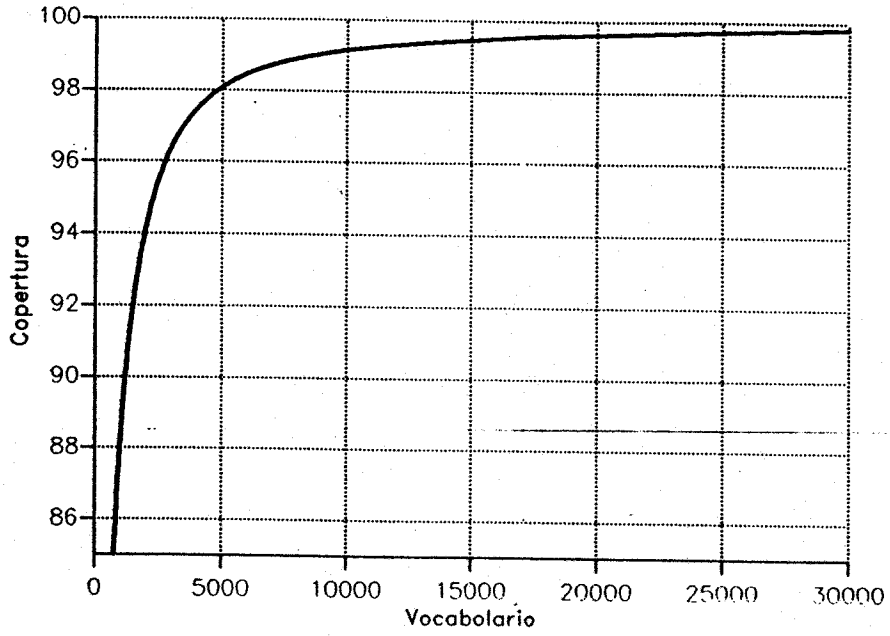

Figura 1. Copertura del corpus OS in funzione della dimensione del vocabolario.

Per questo tipo di lessico, e con la quantita' di dati disponibile, un vocabolario di 5000 parole rappresenta un buon compromesso fra l'esigenza di avere elevati valori di copertura e quella di avere un sufficiente quantitativo di dati per la stima dei parametri del modello del linguaggio come verra' spiegato piu' approfonditamente nel paragrafo successivo.

Allo scopo di ottenere una buona personalizzazione del sistema, si e' scelto di inserire nel vocabolario tutte le 3200 parole distinte individuate nel corpus AP. Il vocabolario e' poi stato completato con altre 1900 parole scelte fra le piu' frequenti del corpus OS e non incluse nelle precedenti 3200.
Come indice per ordinare le parole di OS e' stata usata la media delle frequenze di occorrenza di ciascuna parola in ognuno dei corpora OS_i:

$$\bar{f} = \frac{1}{4} \times \sum_{i=1}^{4} \frac{C_{OSi}(w)}{N_{OSi}} \qquad (3)$$

e non la frequenza assoluta:

$$f = \frac{\sum_{i=1}^{4} C_{OSi}(w)}{\sum_{i=1}^{4} N_{OSi}} \qquad (4)$$

($C_{OSi}(w)$ indica il numero di volte che la parola w appare nel corpus OS_i, mentre N_{OSi} indica la dimensione del corpus OS_i).
Infatti, poiche' i corpora provenienti dai 4 ospedali hanno dimensioni diverse, se si fosse adottato il secondo indice come criterio classificatore, sarebbero state penalizzate le parole provenienti dai corpora di dimensioni minori.
Il dizionario cosi' costruito ha una copertura del 100% sui referti di tipo VI (tutte le parole di VI sono state infatti inserite nel dizionario) e una copertura del 98% sui referti di tipo OS.

Applicazione PA Per estrarre le parole del vocabolario PA, sono stati analizzati i testi prodotti dal una compagnia di assicurazioni. Tali testi sono stati raccolti presso 9 dipartimenti e sono relativi ad argomenti differenti: lettere d'ufficio, stipula di polizze di vario tipo, documenti di diversa natura, ecc.
L'analisi della copertura incrociata fra i 9 corpora divisi per dipartimento, ha mostrato come questi facciano uso di lessici fra loro molto distanti. Si e' pertanto ritenuto opportuno basare il dizionario sui testi provenienti da un solo dipartimento, costruendo cosi' un'applicazione mirata per quel tipo di lessico.
Il corpus selezionato come il piu' adatto per la nostra applicazione ha 1.5 milioni di parole e una buona affinita' con il lessico economico-finanziario (copertura con il dizionario EF pari al 95.2%).
Per la definizione del vocabolario PA, sono state selezionate le 15000 parole piu' frequenti del dizionario EF e a queste sono state aggiunte le 3100 parole piu' frequenti del corpus PA non incluse nelle precedenti 15000.
Il dizionario cosi' costruito ha una copertura del 99% sui testi di tipo PA e una copertura del 95.7% su testi di tipo economico-finanziario.

Costruzione del modello statistico del linguaggio per le diverse applicazioni

Come gia' detto, il modello statistico di linguaggio ha il compito di stimare la probabilita' *a priori* $P(\overline{W})$ che il parlatore voglia pronunciare la sequenza di parole W. Essa puo' essere valutata come segue:

$$P(\overline{W}) = \prod_{i=1}^{N} P(w_i | w_1, \ldots, w_{i-1}) \qquad (5)$$

Se V e' la dimensione del vocabolario, per ogni i esistono V^i *contesti*, ovvero V^i sequenze di parole $w_1, \ldots, w_{i-1}$. E' dunque chiaro che per specificare $P(w_i | w_1, \ldots, w_{i-1})$ sarebbe necessario calcolare e memorizzare V^i grandezze, e cio' non e' ovviamente possibile anche per piccoli valori di i, quando si ha a che fare con vocabolari di migliaia di parole.

Non e' dunque possibile stimare la probabilita' che la parola w_i verra' pronunciata utilizzando per la predizione ***tutte*** le parole pronunciate prima di essa.
D'altra parte, ci si puo' basare sulla considerazione intuitiva che le parole piu' recenti hanno una maggiore influenza su cio' che verra' pronunciato. E' cosi' possibile considerare equivalenti le frasi che terminano con le stesse M parole. Circa il valore da scegliere per M, esso deve risultare da un compromesso tra quantita' di dati da trattare ed efficacia predittiva. Nel nostro caso la scelta e' stata $M = 2$ [10] : si parla in tal caso di modello a *trigrammi* .

Il modello di linguaggio dovra' stimare pertanto $P(w_i|w_{i-1}, w_{i-2})$. Il numero di trigrammi $w_1w_2w_3$ possibili e' V^3 (8×10^{11} se $V = 20.000$). Poiche' l'ordine di grandezza dei testi su cui vengono stimate le statistiche e' in genere di $10^7 \div 10^8$, e' chiaro che, anche tenendo in considerazione i vincoli grammaticali, il numero di trigrammi possibili e' troppo alto per avere una quantita' sufficiente di dati di addestramento per ciascuno di essi. E' allora necessario in generale ricorrere a tecniche di *smoothing* che consistono nel combinare le informazioni fornite dai trigrammi con quelle provenienti dalle distribuzioni di *bigrammi* (coppie di parole) e *unigrammi* (parole isolate).

Nel nostro caso e' stata utilizzata una tecnica di interpolazione delle informazioni fornite dalle varie distribuzioni:

$$P(w_3|w_1w_2) = \lambda_3 \frac{C(w_1w_2w_3)}{C(w_1w_2)} + \lambda_2 \frac{C(w_2w_3)}{C(w_2)} + \lambda_1 \frac{C(w_3)}{N} + \lambda_0 \frac{1}{V} \quad , \tag{6}$$

dove con $C(w_1, \ldots, w_n)$ si indica il numero di occorrenze della stringa $w_1, \ldots, w_n$ nei dati di addestramento del modello. Il quarto termine a secondo membro della (6) e' il cosiddetto termine a *0-grammi*. I coefficienti λ sono tali che

$$\lambda_3 + \lambda_2 + \lambda_1 + \lambda_0 = 1 \tag{7}$$

Essi vengono stimati tramite l'algoritmo di *expectation-maximization (EM)*[11][12][13][14][15] massimizzando la probabilita' di un testo di riferimento (T_{no}) nel testo di addestramento T_1.

In questa stima, un problema da tenere in considerazione e' dato dal fatto che e' conveniente utilizzare piu' di un insieme di coefficienti: cio' permette di variare il *peso* delle distribuzioni a seconda della loro affidabilita' statistica.
Quanto detto si puo' spiegare con l'osservazione che e' ragionevole aspettarsi un valore di λ_3 tanto piu' prossimo a 1 quanto piu' $C(w_1, w_2)$ e' grande. Nel nostro caso sono stati utilizzati fino a trenta insiemi di coefficienti corrispondenti a vari valori di $C(w_1, w_2)$.

L'algoritmo per la stima dei coefficienti λ si puo' schematizzare nei passi seguenti:

1. Scelta di un valore iniziale per ogni λ_{bi} , dove $i = 0, 1, 2, 3$ e b e' l'indice del generico insieme di coefficienti;
2. Calcolo di

$$\hat{\lambda}_{bi} = \frac{count(b, i)}{\sum_{i=0}^{3} count(b, i)} \tag{8}$$

dove

$$count(b, i) = \sum_{k=1}^{N_{no}} \frac{\lambda_{bi} P_i^k C_k}{\sum_{i=0}^{3} \lambda_{bi} P_i^k} \tag{9}$$

Nella (9) N_{no} e C_k sono, rispettivamente, il numero di parole che compongono il testo T_{no} ed il numero di occorrenze del suo $k-esimo$ trigramma. P_i^k , dove $i = 0, 1, 2, 3$ sono i termini di $i-gramma$ del membro di destra della (6).
3. $\lambda_{bi} = \hat{\lambda}_{bi}$;
4. Ritorno al passo 2) fino al raggiungimento della convergenza.

Una misura del tasso di convergenza dell'algoritmo *EM* puo' essere ottenuta calcolando ad ogni iterazione la quantita' $PP = 2^{\tilde{H}}$, dove $\tilde{H}$ e' data da:

$$\tilde{H} = -\frac{1}{N_{HO}} \sum_{i=1}^{N_{HO}} \log_2 P(w_i | w_{i-1} w_{i-2}) \qquad (10)$$

PP e' nota come *perplessita'* ed e' un indice generalmente usato per valutare il potere predittivo di un modello statistico di linguaggio [16]: essa da' il numero di parole del vocabolario considerate equiprobabili dal modello. In assenza del modello, come e' ovvio, $PP = V$.

Nella figura 2 e' mostrato un grafico della perplessita' del testo di riferimento rispetto al numero di iterazioni dell'algoritmo *EM*.

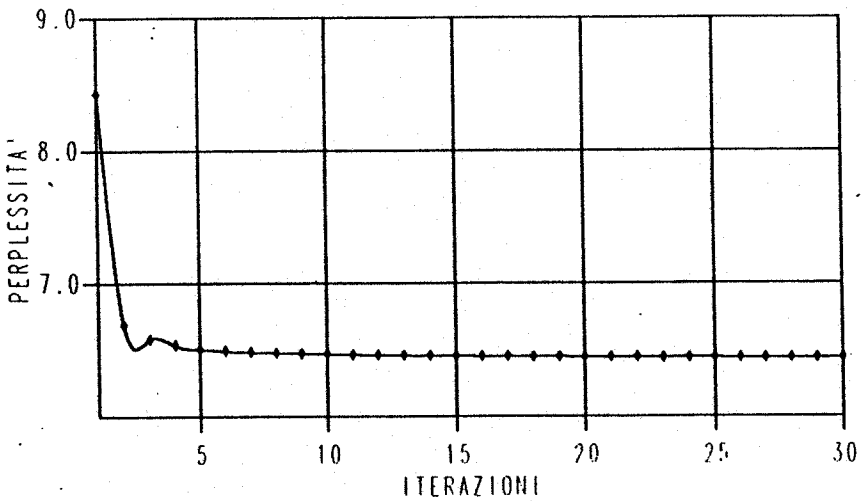

Figura 2. Tasso di convergenza dell'algoritmo EM stimato attraverso la perplessita' misurata su un testo di riferimento.

A differenza di una grammatica, un modello statistico di linguaggio non ha un carattere *prescrittivo*, ma, piuttosto, *descrittivo*: cio' lo rende preferibile per realizzare applicazioni in linguaggio naturale, in cui si fa uso di vocabolari di grandi dimensioni. Il prezzo che si deve pagare e' pero' costituito dalla necessita' di disporre di un cospicuo quantitativo di dati di addestramento. A tale proposito, nel dominio lessicale "Economia e Finanza" sono stati compiuti vari esperimenti addestrando modelli di linguaggio per un dizionario di 20000 parole con diverse quantita' di dati ed eseguendo per ciascun modello misure di perplessita' e di tasso di riconoscimento.

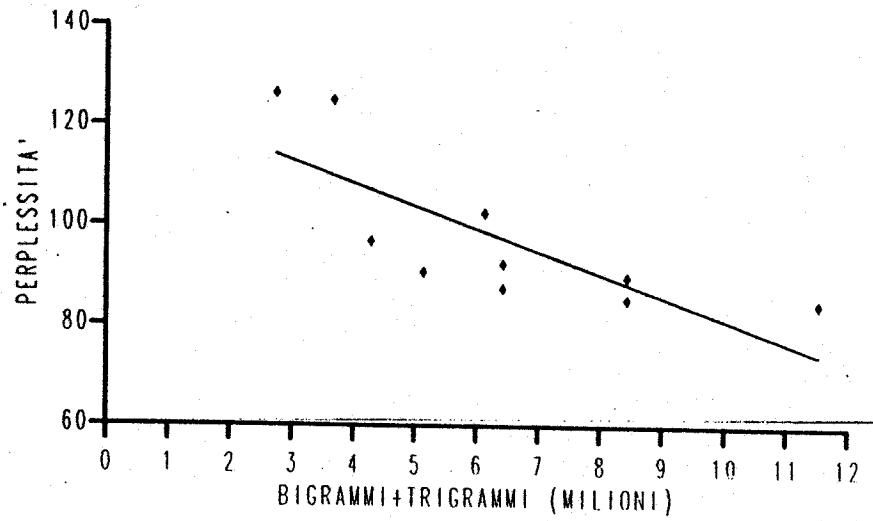

Figura 3. Misure di perplessita' per vari modelli di linguaggio in funzione della quantita' di dati di addestramento.

Nelle figure 3 e 4 sono riportate le prestazioni di vari modelli statistici di linguaggio in funzione della quantita' di dati di addestramento. Nelle figure le linee continue sono un best-fit realizzato con il metodo dei minimi quadrati. Si e' preferito usare come ascissa la somma dei numeri di trigrammi e bigrammi diversi presenti nei dati di addestramento dei modelli in quanto questa grandezza e' quella piu' direttamente legata al potere predittivo di ciascun modello. Complessivamente, sono stati fatti esperimenti con un corpus di addestramento variabile da 21.5 a 107 milioni di parole; in corrispondenza di cio' la perplessita' e' scesa da 126 a 85 ed il tasso di errore dal 4.25% al 3.49% (circa -18%). La perplessita' e' stata misurata utilizzando un testo di circa 43000 parole proveniente dallo stesso dominio linguistico cui appartengono i dati di addestramento dei modelli ma disgiunto da questi. I dati relativi al tasso di riconoscimento sono stati ricavati come media su un campione di 24 parlatori. Per ogni parlatore e' stata eseguita la decodifica di un insieme di frasi pre-registrate per complessive 1043 parole per ogni parlatore.

Utilizzando la tecnica descritta in questo paragrafo, sono stati realizzati modelli statistici di linguaggio per le applicazioni **RR** e **PA** oltre a quello realizzato per il vocabolario economico-finanziario. Nella Tabella seguente sono riportati i parametri piu' significativi per i due modelli di linguaggio.

Tabella 1. Caratteristiche di modelli di linguaggio per due le applicazioni.. I dati di addestramento ed i numeri di bigrammi e trigrammi sono espressi in milioni di parole. Nell'applicazione PA e' stato usato un corpus di 1.5 milioni di parole. di parole tipico del dominio lessicale dell'applicazione. Esso e' stato integrato da un corpus di 40 milioni di parole del dominio EF.

Parametro	RR	PA
Dimensione del vocabolario	5100	18100
Dati di addestramento	4.8	1.5 + 40
Trigrammi diversi	0.62	3.4
Bigrammi diversi	0.19	2.3
Perplessita'	38	18

Risultati

Sono state realizzate due applicazioni per la dettatura di referti radiologici e di polizze di assicurazione. Vengono qui di seguito riportati i risultati di entrambi gli esperimenti. In ambedue i casi si tratta di risultati ancora non definitivi, ma che possono gia' dare un quadro rappresentativo delle prestazioni del sistema.

Dettatura di referti radiologici Quattro medici hanno partecipato all'esperimento. Tutti hanno mostrato notevole adattabilita' al sistema, sia come capacita' di utilizzare le varie funzioni sia per quanto riguarda l'uso della dettatura in **parlato connesso** (inserzione di una brevissima pausa tra le parole). Tutti i partecipanti avevano limitate conoscenze informatiche.: Sono stati dettati 152 referti per un totale di oltre 12000 parole. La copertura del vocabolario relativamente alle parole dettate e'

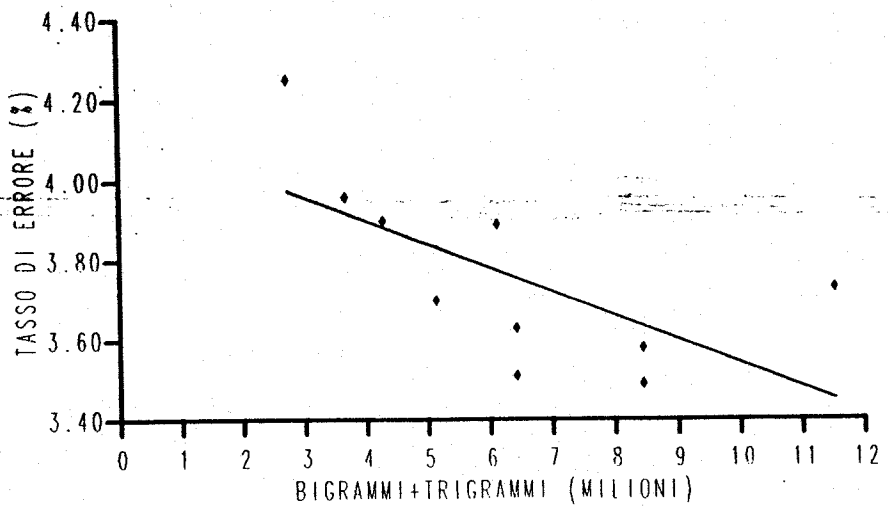

Figura 4. Misura del tasso di riconoscimento per vari modelli di linguaggio in funzione della quantita' dei dati di addestramento.

stata del 98.7%. Nella tabella seguente sono riportati, per ogni parlatore, i risultati complessivi dell'esperimento.

Tabella 2. Risultati dell'esperimento di dettatura di referti radiologici.			
Parlatore	**Referti dettati**	**Tasso errore netto**	**Tasso errore umano**
p1	25	1.7%	1.4%
p2	12	2.0%	1.2%
p3	79	3.5%	0.9%
p4	36	5.0%	3.2%

Il tasso di errore netto si riferisce agli errori effettivamente attribuibili al riconoscitore. Il tasso di errore umano si riferisce ad errori dovuti ad un uso errato della macchina (mancato inserimento delle pause tra le parole, errori nell'uso dei comandi).
Una stima globale del comportamento del sistema si puo' avere sommando al tasso di errore netto la percentuale di errori dovuti all'assenza della parola nel vocabolario che, come visto, e' stata in media dell'1.3%. In tal modo si ottengono dei tassi di riconoscimento variabili tra il 93.7% ed il 97% (che scendono, rispettivamente, al 90.5% ed al 95.6% se si tiene conto anche degli errori umani).

Dettatura di polizze di assicurazione In questo caso sono disponibili risultati su tre parlatori. Sono state dettate oltre 8000 parole con una copertura del vocabolario intorno al 99%. A differenza dell'esperimento precedente, in questo caso i parlatori avevano gia' esperienza nell'uso di sistemi informatici. I risultati dell'esperimento sono riportati nella tabella seguente.

Tabella 3. Risultati dell'esperimento di dettatura di polizze di assicurazione		
Parlatore	**Tasso errore netto**	**Tasso errore umano**
p1	1.9%	1.0%
p2	1.4%	0.5%
p2	2.4%	0.7%

Osservazioni Nonostante i risultati esposti siano da considerare preliminari, possono essere tratte utili indicazioni.
E' da notare come i vari utenti abbiano rapidamente imparato ad usare la macchina per dettare. Cio' si puo' desumere dai bassi valori del tasso di errore umano che, tranne in un caso, si sono sempre mantenuti sotto il 2%.
Il tasso di errore netto e' stato soddisfacente. Il numero di errori nei testi dettati dovuti alla mancanza di parole nel vocabolario e' risultato molto contenuto. Un utile risultato degli esperimenti e' stato quello di raccogliere indicazioni su come ampliare i vocabolari in modo tale da estendere il dominio lessicale delle applicazioni.
Infine, e' da notare come l'applicazione di dettatura dei referti radiologici si sia potuta inserire in un contesto "produttivo" con ottimi risultati. In particolare, sono stati notevolmente abbassati i tempi necessari per la stesura definitiva dei referti ed inoltre il sistema si e' rivelato estremamente utile per gli esami di pronto soccorso notturni e festivi, quando non e' disponibile alcun servizio di dattilografia.

BIBLIOGRAFIA

[1] F. Jelinek, **The Development of an Experimental Discrete Dictation Recognizer,** *Proceedings of IEEE,* vol. 73, no. 11, November 1985, pp. 1616-1624.

[2] P. D'Orta, M. Ferretti, A. Martelli, S. Melecrinis, S. Scarci, G. Volpi, **Large-Vocabulary Speech Recognition: a System for the Italian Language,** *IBM Journal of Research and Development,* Vol. 32, No. 2, March 1988, pp.217-226.

[3] P. Alto, M. Brandetti, M. Ferretti, G. Maltese, S. Scarci, **Esperimenti di dettatura in linguaggio naturale con un riconoscitore per 20000 parole,** *Atti Congresso annuale AICA, Trieste, 4-6 ottobre 1989,* vol. 1, pp.363-369.

[4] P. D'Orta, M. Ferretti, S. Scarci, **Riconoscimento della voce per la dettatura automatica di testi,** *Atti Congresso annuale AICA, Trento, 30 settembre - 2 ottobre 1987,* vol. 2, pp.111-128.

[5] L. R. Bahl et al., **Acoustic Markov Models Used in the Tangora Speech Recognition System,** *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing,* New York, April 1988, Vol. 1, pp. 497-500.

[6] R. Carlson, B. Granstroem, **A Text-to-Speech System Based Entirely on Rules,** *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing,* Philadelphia, PA, April 1976.

[7] D. H. Klatt, **Structure of a Phonological Rule Component for Synthesis-by-Rule Program** *IEEE Trans. on Acoust., Speech and Sig. Proc.,* vol. ASSP-24, no. 5, 1976, pp. 391-398.

[8] T. J. Sejnowski, C. R. Rosenberg, **Parallel Networks that Learn to Pronounce English Text,** *Complex Systems,* 1 (1987), pp. 145-168.

[9] S. Scarci, S. Taraglio, **Automatic Phonetic Transcription for Large-Vocabulary Speech Recognition,** *Speech 88, Seventh FASE Symposium,* Edinburgh, 22-26 August 1988, pp. 771-777.

[10] F. Jelinek, **Self Organized Language modeling for Speech Recognition,** *IBM Internal memo,* February 1986.

[11] A.M. Derouault and B. Merialdo, **Language modeling at the syntatic level,** *Proceedings of the VII International Conference on Pattern Recognition* July 30 - August 2, 1984.

[12] A.M. Derouault and B. Merialdo, **Natural Language Modeling for Phoneme-to-Text Transcription,** *IEEE Transcactions on Pattern Analysis and Machine Intelligence, vol. PAMI-8, no. 6, pp. 742-749,* 1986.

[13] F. Jelinek, R.L. Mercer, **Interpolated Estimation of Markov Source Parameters from Sparse Data,** *Pattern Recognition in Practice,* E.L.Gelsema and L.N. Kanal, Ed., North-Holland, New York, 1980, pp. 381-397.

[14] A.P. Dempster, N.M. Laird and D.B. Rubin, **Maximum Likelihood from Incomplete Data via the EM algorithm,** *J. of Roy. Stat. Soc., no. 1, pp. 1-38,* 1977.

[15] P.F. Brown, **Speech recognition by Statistical Methods,** *IBM internal Seminar,* November 1985.

[16] F. Jelinek, R.L. Mercer, L.R. Bahl and J.K. Baker, **Perplexity - a measure of the difficulty of speech recorgnition tasks,** *Program of 94th Meeting of the Acoustical Society of America J. Acoust. Soc. Am. vol. 62, Suppl. no. 1, p. S63* 1977.